# Circulating cell-free DNA methylation analysis of metastatic prostate cancer

**An-Jui Wu**

Division of Molecular Pathology,
the Institute of Cancer Research

The thesis submitted to the University of London for the Degree of Doctor of Philosophy in the Faculty of Science, 2019

# Preface & Acknowledgement

*"I won't say do not weep; for not all tears are an evil."*

*— J.R.R. Tolkien, The Return of the King*

Four years ago, when I took an one-way flight to London from Taipei, the light of excitement and the darkness of uncertainties casted over me as I was not fully sure what this journey would bring me. It was a bright, crisp autumn day when I landed at London Heathrow, and by that time I had absolutely no idea what was awaiting me. I always have a dream to create and innovate the ways things were done and make inventions that can change the world. I chose the topic of circulating cell-free DNA methylation analysis simply because I believed it would ultimately lead to paradigm-changing discoveries. You all readers will be able to judge on this matter after reading my thesis. In the past 48 months, I took few small courageous steps in research, made some interesting findings, and most importantly still managed to enjoy the science that I was doing. Every beginning has an end, and now it comes to the end of my PhD. I am certain it is just the end of the beginning. It is difficult to summarise my PhD journey in few words or paragraphs, but if I do have to do so, I would like to conclude with a big THANK YOU. A massive thank you goes to everybody who took part in my PhD, and my life in the UK would not be as interesting and rewarding without every one of you.

First of all, I would like to thank my parents, especially my mom. You were really supportive and always there for me. The allegiance and devotion were extremely essential for me throughout my study.

I would like to pay my great gratitude to my supervisors who guided me through all the difficulties in the past four years. Thank you to Prof. Gerhardt Attard for being my primary supervisor and investing huge amount of efforts and capitals on my projects. I would not be able to start and conclude my projects without your support and guidance. Thanks to Dr Daniel Wetterskog for being my day-to-day supervisor. Your experience and attention to details were crucial to my experimental designs and executions. Thank you to Dr Paolo Cremaschi for your assistance and expertise in bio-informatics. Your advice and analyses were key to our breakthrough in the plasma methylome research.

I would like to pay tribute to all the Attard's lab members, especially Dr Karolina Nowakowska, Ania Wingate, Dr Anuradha Jayaram, Lesley Carr, Dr Emily Grist, Dr Mariana Buongermino Perreira, Dr Maria R. Vico, Dr Francesco Pierantoni and Dr Ismail Mazlina. I would like to mention especially Dr Karolina Nowakowska; you are my closest friend in the team, and your warm welcome plus day-to-day support was very important for me to go through all the hurdles.

I would like to mention one my best friend and ex-colleague – Dr Dimitrios Kleftogiannis. Your enthusiasm in data science and cancer biology truly inspired

Masurel, Dr Jasmin Strauss, Dr Marine Anquetil, Catherine Hsu, and Michal Pawelkowicz. I appreciated all the interesting ideas and smart advice you shared with me and it was my great pleasure to have you in the journey.

I would like to express my gratitude and admiration to my funding bodies – Taiwan Ministry of Education, Prostate Cancer Foundation, Cancer Research UK, and the Institute of Cancer Research. Without your funding and support, I would not be able to complete my study and make discoveries. Last but not least, I thank all the participating men and their families who suffered from metastatic prostate cancer and nonetheless gave the gift of participation so that others might benefit.

# Declaration

I declared that the work I presented here has been performed by me (Anjui Wu) unless otherwise acknowledged.

This PhD thesis has led to three publications and one patent at UK IPO ("Cancer Detection Methods", UK application no. 1915469.9), sponsored by Cancer Research UK and UCLB (UCL Business).

- **Wu, A**., et al. (2020). Genome-wide plasma DNA methylation features of metastatic prostate cancer. The Journal of Clinical Investigation (*accepted*).
- **Wu, A**., et al. (2019). Pan-genome cfDNA methylation analysis of metastatic prostate cancer. Annals of Oncology (2019) 30 (suppl_7): vii1-vii35. 10.1093/annonc/mdz238
- **Wu, A**., and Attard, G. (2019). Plasma DNA Analysis in Prostate Cancer: Opportunities for Improving Clinical Management. Clin Chem *65*, 100-107.

# Academic accomplishments

## Conferences

- Proffered oral presentation – 'pan-genome cfDNA analysis of metastatic prostate cancer' in Molecular Analysis for Personalised Therapy (MAP), London UK, 2019
- Commercial patent application - multiple purposes of cfDNA methylation analysis in detection, screening, monitoring, staging, classification and/or prognostication of prostate cancer
- Poster presentation in Prostate Cancer Foundation (PCF) Annual Science Retreat, Carlsbad, San Diego USA, 2019
- Poster presentation in EMBL-EBI Cancer Genomics Conference, Heidelberg Germany, 2019
- Poster presentation in UCL Cancer Institute Annual Conference, London UK, 2018
- Poster presentation in Institute of Cancer Research Annual Conference, Egham UK, 2018
- Poster presentation in Cancer Evolution Conference, Hinxton UK, 2018
- Oral presentation in Urological Congress, Meldola Italy, 2017

## Honours

Goodenough College Studentship
2019

University of London Conference Fund
2017

cfDNA Conference Merit Award
2017

Wellcome Trust Conference Funds (two times)
2017, 2018

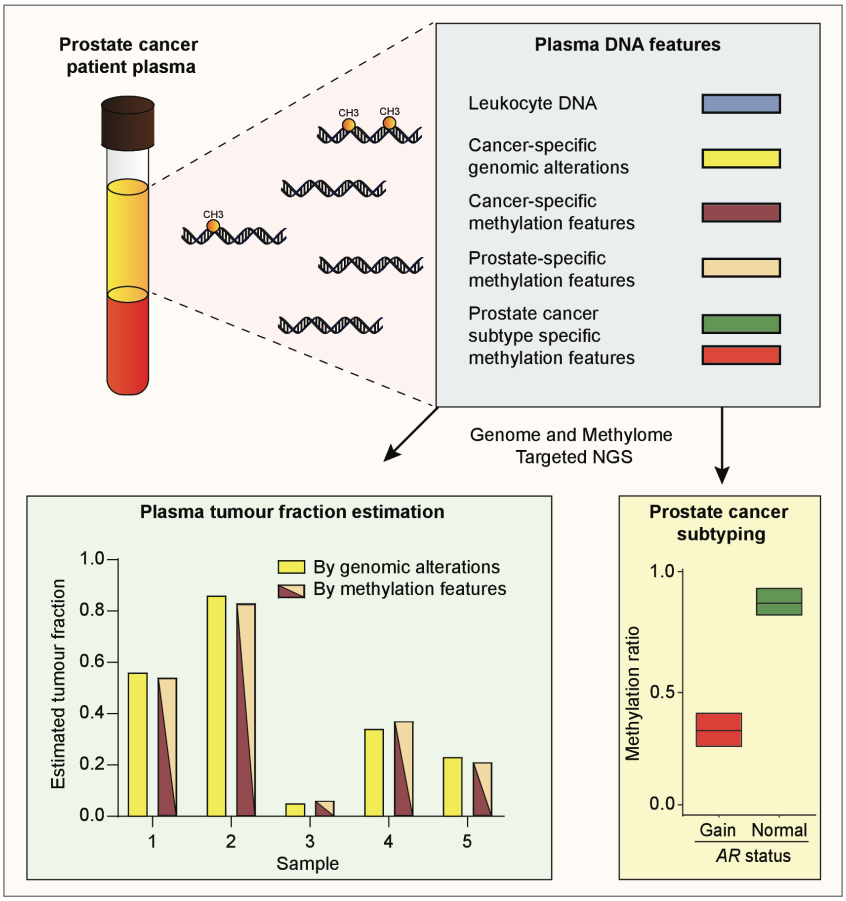
CRUK Young Investigator Travel Award
2016

Institute of Cancer Research Scholarship
2017

Taiwan Young Elite Scholarship
2015

# Table of Contents

# Graphic Summary



# 1 Chapter 1. Introduction

## 1.1 Overview

Prostate cancer is one of the leading causes of death for men in the UK (Cancer Research UK, https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer, accessed 10/2019). It is a highly heterogeneous disease with variable clinical outcome and response to treatment. The screening of early prostate cancer relies on circulating prostate-specific antigen (PSA), although its use in national screening remains controversial. Men at high risk of prostate cancer are subjected to trans-rectal prostate biopsy, the results of which are used in conjunction with PSA level and whole-body imaging for patient stratification. The locally-advanced non-metastatic intermediate to high risk group would accordingly receive curative local treatment such as prostatectomy or radiotherapy. Metastatic disease is treated with long-term anti-androgen therapy (ADT, i.e. luteinizing hormone releasing hormone, or LHRH, agonist or antagonist) and usually progress to castration-resistant lethal disease. Recent evidences have showed that docetaxel chemotherapy or novel androgen receptor or AR-axis targeting agents such as abiraterone [1,2] or enzalutamide [3] added at start of ADT improve survival and delay time to castration-resistance.

Multiple molecular characterisation studies of localised and metastatic prostate cancer have been done and have started to dissect tumour carcinogenesis, evolution and underlying treatment resistance. Despite recent advances, there

are still many challenges remaining. For example, the molecular aberrations that drive aggressive disease and drug resistance are largely unknown.

Plasma cell-free DNA analysis has the potential for cancer detection at an earlier stage or at relapse before cancer becomes radiologically detectable. Furthermore, it can be used to identify emerging resistant tumour clones in metastatic disease. Analysis of plasma DNA somatic point mutations or copy number alterations through liquid biopsy has the potential to inform treatment decisions in cancer patients with a range of tumour types [4,5]. For example in lung cancer, plasma-based testing to screen for epidermal growth factor receptor (EGFR) point mutation T790M is a powerful method to select patients eligible for third generation EGFR inhibitors [6] [7]. Several studies have shown that plasma DNA is representative of clinically relevant metastases[8]. Importantly, plasma DNA could share driver DNA alterations with matched metastatic tissue samples but some rare mutations may be private to plasma samples.

Additional to genomic information, plasma DNA also contains methylation information that can be concurrently extracted. Methylation status is tissue-specific and can be used to interrogate cellular components and quantitate tissue composition and tumour origin and potentially changes in methylation could underlie treatment resistance that could be monitored using plasma [9,10]. Several studies to date have used methylation information from plasma DNA for the early detection of cancer and to identify cancer tissue of origin [11-19], but the plasma DNA

methylome has not been as extensively studied and characterised in patients with metastatic cancer.

## 1.2 Prostate cancer diagnosis and management

Prostate cancer is the most prevalent cancer in men in the United Kingdom and Europe. It is estimated that more than 2.6 million men are diagnosed with prostate cancer worldwide each year, resulting in over 900,000 deaths[20,21].

Prostate cancer is a very heterogeneous disease characterised by variable clinical outcome. Measurement of serum PSA levels is used for the early detection of potential prostate cancer and if found elevated could lead to a diagnostic biopsy in the form of a trans-rectal ultrasound-guided biopsy (TRUS-biopsy). However, even with an elevated PSA, many biopsies return negative[22]. Biopsies can also yield false negatives, particularly when not guided by imaging[23]. There has therefore been a paradigm shift towards more accurate imaging-guided biopsies. Researchers in the PROMIS study (NCT01292291) have demonstrated the feasibility and benefit of Multi-Parametric Magnetic Resonance Imaging (MP-MRI) for patient triage. The results showed that MP-MRI guided biopsy can reduce the number of unnecessary biopsies by over 25%, while also reducing the over-diagnosis of clinically insignificant diseases. The multi-centre, randomised PRECISION study showed that an MRI-targeted biopsy was able to pick up more clinically-significant tumours as compared with standard biopsy, and fewer men in MRI-guided group were diagnosed with clinically insignificant tumour than these receiving standard biopsy [24].

Risk stratification of prostate cancer is based on TNM score, Gleason score and PSA values (National Comprehensive Cancer Network, or NCCN, guideline for

prostate cancer, Inc. 2018). For localised low-risk prostate cancer, the chance of metastasis is low, and treatment options would be based on life expectancy and quality of life. In fact most localised low-risk prostate cancers remain indolent and do not require active treatment. Patients with localized intermediate-risk prostate cancer can consider radical local treatment such as prostatectomy, radiotherapy or brachytherapy with the potential addition of concurrent chemotherapy or anti-androgen therapy (ADT). Patients with locally advanced, metastatic or high-risk disease are strongly advised to receive radical treatment followed by anti-androgen therapy. However, most of these patients recur and androgen receptor (AR) targeting therapies with or without systemic chemotherapies are then introduced to control disease progression.

For patients with metastatic diseases, continuous ADT was first introduced in 1941 by Nobel Laureate winner Dr Charles Huggins, and continuous ADT on its own remains the standard-of-care. Despite initial response rates of over 80%, prostate cancer eventually progresses and becomes resistant to ADT; this is referred to as castration-resistant prostate cancer (CRPC), as opposed to castration-sensitive prostate cancer which still responds to ADT. It was hypothesized that early introduction of chemotherapy or next-generation AR-targeting agents such as abiraterone or enzalutamide could control early aggressive tumour clones or existing clones resistant to ADT. Docetaxel (Taxotere), which works by binding to microtubules and preventing cell mitotic division, was first introduced to treat advanced prostate cancer in 2004 [25]. In 2015, Androgen Ablation with or without Chemotherapy in Treating Patients with Metastatic Prostate Cancer (CHAARTED

trial, NCT00309985)[26] showed that adding the chemotherapeutic agent docetaxel along with anti-androgen therapy (i.e. LHRH agonist or antagonist) conferred significant survival benefit. The initial report from a multi-centre, multi-stage, multi-arm prospective clinical trial Systemic Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy (STAMPEDE trial, NCT00268476) also showed similar survival benefits of adding docetaxel to ADT [27].

Abiraterone acetate (or abiraterone) [28-30], an AR-targeting agent which inhibits CYP17 (a key enzyme in the androgen biosynthesis pathway) and weakly antagonises the androgen receptor, was first introduced to treat CRPC with [31] or without prior docetaxel treatment [32] based on the COU-AA-302 and COU-AA-301 study, respectively. Enzalutamide (MDV3100), designed to inhibit androgen receptor function, has also improved survival in CRPC patients with or without previous chemotherapy [33,34].

In 2017, clinical readouts from the STAMPEDE trial and the LATITUDE trial[1] showed that combination of ADT and abiraterone improved patient survival [2]. Similar to STAMPEDE and LATITUDE, the ENZAMET study (NCT02446405) studied the possibility of adding enzalutamide to the first-line standard-of-care treatment on CSPC patients with or without early docetaxel. The result showed that early introduction of enzalutamide improved progression-free survival (PFS) and overall survival (OS) [3]. Moreover, apalutamide, another novel AR-targeting agent which directly binds to the androgen receptor ligand binding domain [35], proved to

increase progression-free survival and overall survival in metastatic CSPC patients (TITAN trial, NCT02489318) [36].

The positive results of TITAN, STAMPEDE, CHAARTED, ENZAMET and LATITUDE trials, introduced many new clinical questions. For example, the mechanisms of drug resistance remain unclear for most treatments. There is no clear molecularly-defined risk-stratification strategy that allows clinicians to identify patients with higher risk of relapse, residual disease, or with poor prognosis who may benefit from the combination therapies, while patients with lower to intermediate risk of relapse may omit from additional treatments. Moreover, there are no clinically-accredited biomarkers to match a patient to the treatment most likely to give the greatest benefit, track treatment responses or identify early resistance tumour clones.

## 1.3 Molecular characterisation of prostate cancer

### 1.3.1 Genomic features of hormone sensitive prostate cancer

Prostate cancer genomic studies have identified multiple chromosomal changes occurring early in prostate carcinogenesis. Structural re-arrangements, involving *ETS* gene fusions, *PTEN* deletion, or *NKX3*.1 deletion, were commonly described, and these early events have also been associated with more aggressive forms of the disease. Studies of localised prostate cancer have revealed massive, complex genomic re-arrangements which may occur inter-dependently [37]. The results from whole genome sequencing of normal prostate tissues and cancerous tissues suggest that the complex and lengthy series of rearrangements may be due to "chromoplexy", a phenomenon that describes inter- or intra-chromosomal large DNA fragment re-assembly. More genomic rearrangements were also significantly linked with higher histological grades. Fraser et al. performed comprehensive genomic profiling of localised, intermediate-risk to high-risk prostate tumours and further confirmed that focal genomic instability, including chromotripsis or katageis, was present in over 20% of the tumours [38].

### 1.3.2 Genomic features of castration-resistant prostate cancer

As the disease progresses, it becomes metastatic and resistant to anti-androgen therapies and the disease at this stage is called CRPC. Genes such as *AR*, *PTEN*, *TP53* are frequently altered in mCRPC [39,40], and large studies have concluded that five major signalling pathways are commonly altered in mCRPC - *AR*, *PI3K*, *WNT* signalling, DNA repair, and cell cycle. However, there are still a fraction of tumours

without detectable genomic aberrations using the pre-designed panel [41] or with aberrant changes of uncertain clinical significance. A deep whole genome sequencing projects in 100 patients with mCRPC aimed to look closer into genomic structural variants that may disrupt the function of key tumour driver genes, or tumour suppressor genes. The findings suggested that DNA repair genes defects were associated with an increasing number of structural variants[15].

### 1.3.3 Epigenetic features of prostate cancer

In addition to genomic changes, epigenetic changes are indicators of aging and environment exposures and could also lead to gene instability and subsequent genomic mutations[42] [43] [44].

Recent studies have shown that epigenetic changes remain stable during the cell cycle, while some *in vitro* models and animal models showed that epigenetic changes can be plastic and reversible [45] [46] [47]. Given the potential of plasticity, tumour epigenetic variation has implications for carcinogenesis, and could also lead to subsequent genomic mutations. For example, some epigenetic changes can silence tumour suppression genes, and lead to uncontrolled cell replication, and subsequently result in numerous genetic mutations [48].

DNA cytosine methylation, also called DNA methylation or CpG methylation, is currently the only DNA epigenetic modification that can be effectively extracted and quantified [49]. CpG methylation plays an important part in multiple biological processes by interacting with specific methyl-CpG binding proteins (MBDs), a key messenger to other transcriptional regulators which result in histone modification, chromatin re-arrangement, and differential gene expressions [50] [51]. Some DNA methylation is believed to remain constant in tumour clones, while some methylation consequences may be later events and result in more malignant forms of cancer [52]. Therefore, DNA methylation signatures have been hypothesized to be an important indicator for both early carcinogenesis and advanced tumour progression with poor clinical phenotype. Most importantly,

DNA methylation events can give us insights into tumour clonal dynamic changes and evolution.

The study of methylation patterns in prostate cancer can be dated back to more than 20 years ago [53], and recurrent methylation at genes such as *GSTP1* [54 55] and *RASSF1A* [56] have been tested in multiple independent studies using different assay designs. For these genes the methylation status has been found to be maintained across different disease stages. A private company MDxHealth developed a tissue methylation-based assay (ConfirmMDx) using three commonly methylated genes - *GSTP1*, *APC*, and *RASSF1* to help address the risk of prostate biopsy sampling errors in men with undetected prostate cancer. Moreover, a report from high-risk metastatic prostate cancer tissues indicated a wide inter-individual heterogeneity, while methylation alterations within the same individual were maintained across all metastases [57]. Brocks et al. profiled both methylation and copy number aberrations and mapped the tumour evolutionary processes and found that the evolution patterns of methylation and copy number changes were consistent, suggesting that genomic and epigenetic variations continued to evolve with the treatment [58]. In addition, metastasis-specific DNA methylation was found to occur in cis-regulatory elements such as AR-bound enhancer domains. These findings suggest that CpG methylation patterns are dynamic in prostate cancer and could potentially complement genomic data for better molecular characterisation of the disease.

Several studies have suggested that DNA methylation status can be disrupted due to genomic aberrations, most notably the *ETS* gene fusion [59,60]. The *ETS* fusion event (most commonly *TMPRSS2-ERG*) is present in more than 30% of early stage prostate cancer but its association with clinical outcome remained inconclusive. When comparing *ETS*-fusion positive and negative tumours a methylation difference was observed where the global methylation patterns in *ETS*-negative cancer were closer to that of the pre-cancerous lesion. A seminal integrative study from the cancer genome atlas (TCGA) also indicated *ERG*-fusions belong to a distinct methylation phenotype [61]. This phenomenon suggested that the binding of a transcription factor can interrupt DNA methylation status and was consistent with the expression profiling study that showed key molecular between *ERG*-fusion and non-*ERG*-fusion tumours [62]. This finding also echoed a previous study stating that transcription occupancy could play a protective role in limiting the spread of DNA methylation into adjacent CpG islands [63]. Thus quantification of the methylation status in prostate cancer related to transcription factor binding can be a proxy of measuring the complex transcriptomic regulatory impacts such as *AR* signalling [58].

Higher coverage sequencing data of benign prostate hypertrophy (BPH), localised prostate cancer and CPRC are still relatively limited. Lin et al. first interrogated paired BPH and localised prostate cancer tissues by enhanced reduced representative bisulfite sequencing (eRRBS) [64]. In this study, a high degree of methylation heterogeneity was evident as some tumours showed more differentially hypermethylated regions and some showed more differentially

hypomethylated regions. The study also described increased methylation in CRPC tissues and identify 13 CpG islands that had persistently increasing methylation levels from benign tissue to localised prostate cancer and from localised cancer to CRPC. Similar results from TCGA research consortium also described global hypermethylation in primary prostate cancer as compared to the normal control especially in tumours with *IDH1/2* mutations [61].

Furthermore, DNA methylation patterns in metastatic, treatment-resistant prostate cancer have been investigated for the identification of aggressive clinical phenotypes. DNA methylation patterns alone were able to classify AR-independent mCRPC-neuro-endocrine tumours from mCRPC-adenocarcinomas, and this preliminary report shows that DNA methylation information can facilitate our understanding of the cause underlying treatment resistance [52].

More recently a study on >100 intermediate risk prostate cancer tissues sought to find a better patient stratification and flag the high risk, non-indolent, localised intermediate-risk prostate cancer [38]. Six out of nine prognostic biomarkers were methylation based suggesting that integration of different sources of molecular information could improve the accuracy in detecting aggressive prostate cancers.

## 1.4 Introduction to circulating cell free DNA in plasma

The understanding of prostate cancer molecular characteristics is crucial for translational research as well as the design, implementation and interpretation of NGS based testing in the clinical setting. Currently, most clinical milestones, risk stratification and disease relapse are not based on molecular characteristics, and that fact introduces an opportunity for implementing molecular-based testing for better patient management. For example, genomic alterations such as *PTEN* loss, *ERG*-fusion, *AR* copy number gain has been seen to associate with worse clinical outcome and can be tested for prognostication. As mentioned earlier, integrative studies combining genomic, methylation, and transcriptomic data have identified nine prognostic biomarkers associating with non-indolent prostate cancer[38]. Deep sequencing in metastatic prostate cancer has also revealed novel structural variants that disrupted gene function, and tandem replication that may promote carcinogenesis[65]. However, one major clinical challenge is to effectively and repeatedly collect tumour tissues from multiple metastatic sites. In metastatic prostate cancer (mPC), more than 90% of metastatic disease locate in the bone, a fact further complicating tissue collection. Hence, the introduction of emerging technology to acquire molecular information from metastatic sites is of urgent clinical need.

Circulating cell free DNA (cfDNA) is fragmented, extracellular nucleic material present in the circulation, and firstly introduced clinically to detect circulating fetal DNA in pregnant women [66]. It also has the potential to characterise tumour in

cancer patients [4]. The origin of cfDNA is not entirely clear, though it is widely believed that the DNA is released from dead cells (such as white blood cells, muscle cells and other tissues) through either apoptotic or necrotic pathways. The size of cfDNA from healthy volunteer has a distinct peak at 166bps, corresponding to the size of DNA wrapped around a single nucleosome. However, the size of cfDNA may vary due to tissue-specific nucleosome occupancy and other unknown factors. For example, in cfDNA derived from patients with hepatoma was shorter than cfDNA from healthy individuals [67].

In cancer patients, tumour cell-derived circulating cell free, or circulating tumour DNA (ctDNA), can be quantitated and characterised. In the past few years, progress in the ctDNA field have given us a clear picture of ctDNA biology and avenues for clinical translation. The ctDNA amount varies depending on tumour type and stage of disease [68]. Generally, ctDNA only make up a very low proportion of total circulating cell free DNA. Thus, in order to detect and accurately quantify ctDNA in plasma, very sensitive and specific methods are required. Given the variable ratio of normal to cancer plasma DNA the first most crucial step is to distinguish DNA of normal tissues from DNA released from cancer cells.

Multiple strategies have been proposed to estimate the ctDNA fraction. Conceptually, to estimate ctDNA fraction, one could track a genomic change that occurs early in carcinogenesis (pre-branching) and therefore present in every cancer cell in that individual. Indeed, in several cancers, the allelic frequency of common and recurrent hot-spot point mutations has been used to track ctDNA [69]

[70]. Proof-of-concept analyses in metastatic breast cancer have used structural variants or somatic mutations identified in tumour tissue, and droplet digital PCR or amplicon-based targeted deep sequencing to quantify circulating tumour DNA levels [69]. This could be further optimized and personalized to track patient-specific mutations identified by multi-regional sequencing [71]. However, prostate cancer does not have commonly recurrent, clonal point mutations and thus requires a broader approach. One strategy is to quantitate a panel of genomic changes that have occurred at an early stage of prostate cancer and if truncal would be present in all metastasizing cells. Two such events that could be used to track tumour content in prostate cancer are mono-allelic deletions associated with ETS gene family rearrangements (primarily involving the oncogenes ERG or ETV1 that fuse with an androgen-regulated promoter) and NKX3.1 deletion on chromosome 8p, strongly linked with prostate cancer development. Either alteration occurs in more than 50% of advanced prostate cancer patients, and has been shown to be clonal in mCRPC [37] [72]. As coverage estimations for quantitating mono-allelic deletions are unreliable in plasma samples with low ctDNA fractions, alternative approaches such as leveraging information of germline heterozygous SNPs could be used to identify the tumour reads harbouring the deletions [73]. In short, the allelic frequency (AF) across heterozygous SNPs should be around 50%; however, in the tumours with deletion events, the AF distribution of all heterozygous SNPs across the whole region would shift. And the magnitude of this shift can be used to estimate ctDNA fraction.

Another approach is to estimate tumour fraction by using variant allele frequency (VAF) from mutation calls in whole exome or targeted next-generation sequencing [40]. This approach requires adjustment for loss of heterozygosity (LOH) for every mutation call, or a conservative assumption that LOH co-occurs with all mutations. It should be noted that this could under-estimate tumour fraction if LOH is assumed when it is not present, or conversely over-estimate if a mutation is in an amplified region. The accuracy of using deletions or mutations to quantify tumour fraction will be dependent on that aberration being present in all, or at least the majority, of clones in circulation. An emerging clone with aberrations not included in the NGS data could be missed. A third approach is to use the magnitude of genome-wide copy number aberrations to estimate tumour fraction which could be especially suited for very advanced prostate cancer [74]. ichorCNA, a software applicable to shallow whole genome sequencing (WGS), estimates tumour ploidy and tumour fraction. This could represent a very economical approach that could be widely implemented across plasma samples but may not be amenable to detect tumour fractions below 8-10%. However, at the very least it could serve as a triage to select samples with higher tumour fraction applicable for further analysis [75] [76].

In addition to the genomic information carried with cfDNA it also bears epigenomic marks [77,78]. Epigenetic features, such as DNA methylation and nucleosome positioning, are tissue-specific and now can be effectively profiled in cfDNA. It is generally hypothesized that the majority of tissue-specific methylation patterns are maintained and thus could serve as a stable and reliable biomarker

for diagnosis or detection of cancer, especially for tumours with unknown origin. There are some challenges of profiling circulating plasma methylome. First methylation changes are dynamic and accumulate due to the natural aging processes. It remains unclear how the methylation patterns change over time, and thus it is difficult to find a proper control for tissue methylation patterns subject to signal deconvolution. Second, there are over 20 million CpG sites across whole genome that have been described to be methylated. The high dimensional methylation features can be hard to both process and interpret. Third, the current technology to profile methylation requires a key step called bisulfite conversion, which convert un-methylated cytosines of CpG dinucleotides to uracils or UpG dinucleotides while methylated CpG site remains the unchanged. This step tends to break double-stranded DNA into smaller, partially single-stranded nucleic acids, which make the down-stream NGS workflow more challenging.

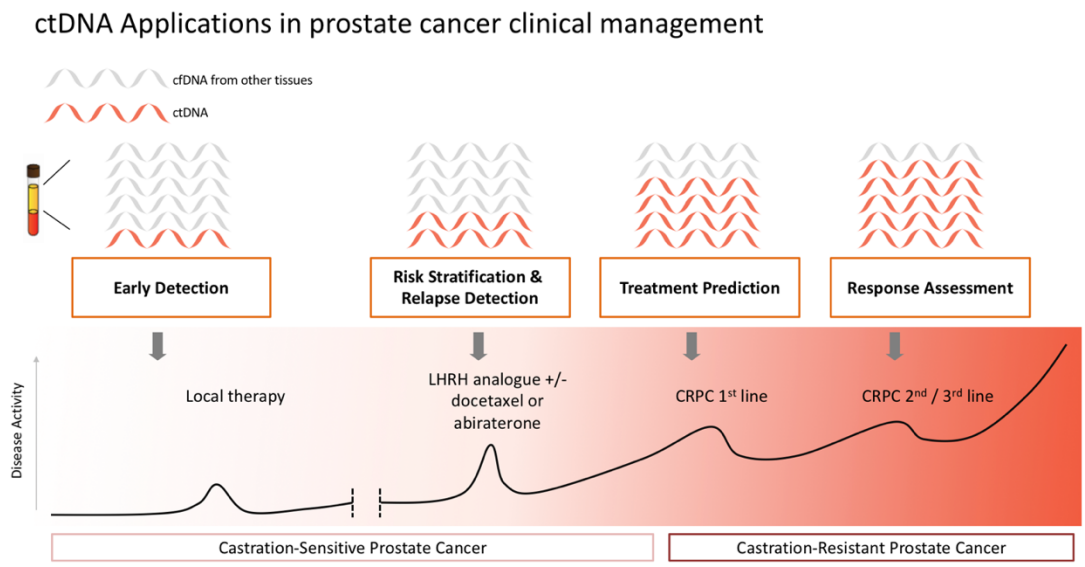
Multiple approaches have been tried to quantify methylation patterns for cancer diagnostic or prognostic purposes. Using a focused approach Xu et.al studied a few informative CpG sites followed by deep sequencing on plasma collected from hepatocellular carcinoma patients. The informative CpG sites were selected based on tissue studies and the rationale was to choose differentially methylated sites between cancerous tissues and normal liver tissues. The collective score of methylation ratio across all selected CpG sites could be used to measure tumour fraction and served as a prognostic biomarker [79]. Apart from the more focused approaches, pan-genome methylation features can also be effectively extracted using window strategies [12-14] and leverages the fact that methylation status of

CpGs across the CpG islands remain consistently hyper- or hypo-methylated. In brief, the pan-genome CpGs were divided into segments and different measurements or simulation were employed to represent the methylation level or heterogeneity. For example, CancerLocator, a probabilistic model for predicting ctDNA burden, used distribution of all methylation ratios within each CpG cluster to simulate methylation level of the cluster [14]. CancerDetector, another software based on Bayesian's and Bernoulli's model, can predict the tissue origin of every single sequencing read [80]. Taken together this opens up more possibilities for future ctDNA applications, especially for early cancer detection, as it allows tracking of tumour-tissue of origin in patients with much improved sensitivity.

## 1.5 Analysis of plasma circulating tumor DNA in prostate cancer

Although ctDNA analysis has the potential to change clinical practice, there is currently no plasma-based test using circulating nucleic acids information to assist clinical decisions and management of patients with prostate cancer. Below, the potential clinical applications of plasma circulating tumour DNA analysis across different stages of the disease is outlined. (**Figure 1.5.1.**).

Figure 1.5.1.



### 1.5.1 Early detection of prostate cancer

Cancer detection and diagnosis at a pre-symptomatic stage could radically improve cancer mortality rates but remains challenging. Improved screening of men for prostate cancer will have major public health benefits – current practices using prostate specific antigen (PSA) result in over-diagnosis of non-lethal disease and over-treatment of several thousands of men every year [81]. Diffusion-weighted pelvic magnetic resonance imaging (MRI) and targeted screening of high-risk men (familial history, germline genetic alterations) are strategies being explored, in combination with PSA, to minimize false positive detection rate [24 82].

The major challenges for a plasma DNA test in this setting are as follows: balancing high specificity and sensitivity in detecting plasma tumour DNA, low ctDNA abundance, and the lack of prior information on the unique molecular features of each individual tumour. In general, cancer screening needs to identify the tumour tissue-of-origin in order to inform clinicians to make actionable plans. Different tumours harbour distinct methylation features, and most changes are tissue-specific. 'CancerLocator' uses methylation status from low-pass whole genome bisulfite sequencing on plasma DNA from lung, breast and colorectal patients and healthy volunteers to build the tumour classifier in order to diagnose cancer with unknown origin [14]. Similar approaches using a customised targeted panel to capture informative CpG sites of hepatocellular carcinoma also showed promising results for cancer detection in patients with liver diseases [79]. Targeted error correction sequencing (TEC-seq) was developed to address the technical hurdle of rare genetic alteration detection without prior tumour information [83]. An ongoing

prospective, multi-centre trial (ClinicalTrial.gov Identifier: NCT02889978) commercially-sponsored by GRAIL, aims to systemically tackle the challenges of early diagnosis by large-scale, multi-centre plasma collection and centralized analysis using NGS-based approaches. Tests for early cancer detection, especially of non-indolent aggressive disease, will need to minimize over-treatment and balance the risks of unnecessary anguish for men who do not require further treatment. This test could be targeted at specific groups, for example based on germline risk factors.

### 1.5.2 Risk stratification and detection of minimal residual disease (MRD) and relapse

Detection of ctDNA shortly after surgical resection or radiotherapy treatment to the primary could be used to stratify patients who require additional systemic treatment. The feasibility of this has been shown in multiple cancer types, including breast, colorectal, and lung tumours. These studies suggest that ctDNA detected shortly after surgery more sensitively predicts tumour relapse than currently used clinicopathological parameters. The risk for relapse in ctDNA positive compared to ctDNA negative patients has been reported as greater than six-fold in multiple studies across different tumour types [71] [70] [84]. This could have important utility in prostate cancer where the risk of relapse is highly variable and could allow selection of adjuvant systemic treatment for the relatively low proportion of patients who would derive maximum benefit. A number of randomized clinical trials in this setting are collecting plasma to evaluate the relationship of ctDNA with treatment response and long-term benefit (examples: ClinicalTrial.gov Identifier: NCT01411332, and NCT01411345). Similarly, analysis of sequential samples from men in follow-up could detect early relapse and initiate life prolonging treatments. In these settings plasma DNA analysis would have to improve on, alone or in combination on serum PSA readings.

### 1.5.3 Prediction of treatment outcome and response assessment in metastatic disease

The first plasma-based test to receive approval from the regulatory authorities for clinical use in cancer patients was the Cobas (Roche™) EGFR Mutation test used to identify EGFR mutations, exon 19 deletion or exon 20 insertions for the selection of patients with metastatic non-small cell lung cancer that stand to benefit from EGFR-targeted therapy.

In mCRPC, mismatch repair (MMR)deficiency occurs in <2% of patient and given that immunotherapy has shown increased efficacy and PD1 blockade has received regulatory approval for use in this molecularly-defined subgroup of patients, there is a potential benefit to test for MMR gene defects in mCRPC patients [85]. DNA repair gene alterations are more common than MMR defects, occurring up to 20% of mCRPC patients. Ongoing trials are selecting mCRPC patients with an underlying germline or somatic DNA repair defect for treatment with agents targeting DNA repair mechanisms, most notably PARP inhibitors [86]. The majority of trials are utilizing archival formalin fixed paraffin embedded (FFPE) tissue or a fresh tissue biopsy for patient selection. Major efforts are underway to concurrently develop a ctDNA-based test. The main challenge remains the accurate detection of mono-allelic (in combination with pathogenic deactivating mutations) and bi-allelic deletions in ctDNA with a highly variable and often low (<0.1) tumour-to-normal fraction.

As CRPC metastases primarily involve bone, quantitative imaging assessment of response or early progression is challenging. Serum PSA is often used in clinical practice to guide decisions on continuing or stopping treatment for disease progression. However, PSA expression is exquisitely androgen-regulated and absolute levels and changes may not entirely reflect disease behaviour and in fact PSA has not met the requirements for a surrogate biomarker of overall survival [87]. Circulating tumour cell dynamics have been shown to strongly associate with treatment benefit across multiple therapeutic strategies. Comprehensive evaluation of CTC changes before and after the treatment indicated that a drop in CTC number in week 13 was strongly linked with prolonged survival [87]. These results are encouraging for liquid biopsy assessment in this setting but the absence of and costs for detection of CTC could limit this application in earlier disease states. Also, ctDNA change in metastatic breast cancer reflective of treatment response had superior sensitivity to CTC and CA15-3 [69]. Preliminary data in mCRPC has indicated that plasma DNA change in sequential plasma samples from mCRPC reflects treatment response [88]. Future studies are required to confirm these findings.

## 1.6 Summary

Clearly, there are a lot of promising results for plasma cell-free DNA analysis changing the clinical management of prostate cancer. Although multiple truncal and sub-clonal circulating genomic events have been quantified and studied, circulating methylation signatures of prostate cancer have yet to be explored. The major goals of my study are to go beyond the genome-centric panorama, to characterise circulating cell-free DNA methylome derived from mCRPC patients, and to identify circulating cancer-specific methylation features. Specifically, I aim to utilise circulating tumour methylation signatures to track plasma circulating tumour DNA fraction, to identify methylation events associated with poor clinical outcomes, to make discovery of emerging resistant clones, and ultimately, to develop an ultrasensitive assay for detection of high-risk diseases at an earlier stage. The mCRPC methylome data generated from this study would also be beneficial for the research community.

# 2 Chapter 2. Materials & Methods

## 2.1 Clinical sample selection & study design

In keeping with this being a discovery analysis in the roadmap to development of a methylation-based biomarker, I selected plasma and tumour samples with sufficient plasma tumour DNA with variable tumour fractions in order to better profile the plasma methylome derived from metastatic prostate cancer. All plasma samples had been subjected to targeted genomic next generation sequencing (NGS) or higher coverage whole genome sequencing to molecularly define tumour fraction (see **section 2.6.1.**).

Plasma samples had been collected within 30 days of treatment initiation and at progression in three biomarker studies, separately approved by the Istituto Scientifico Romagnolo per lo Studio e la Cura dei Tumori (IRST), Meldola, Italy (REC 2192/2013), Royal Marsden, London, UK (REC 04/Q0801/6) and in the PREMIERE trial (EudraCT: 2014-003192-28, NCT02288936) that was sponsored and conducted by the Spanish Genito-Urinary oncology Group (SOGUG) (**Table 2.1.1.**). All patients provided written informed consent for these analyses. These cohorts have been described previously [89]. Briefly, patients needed to have histologically or biochemically confirmed prostate adenocarcinoma and be starting abiraterone or enzalutamide for progressive mCRPC. Patients were required to receive abiraterone or enzalutamide until disease progression as defined by at least two of the following: a rise in PSA, worsening symptoms, or radiological progression

defined as progression in soft-tissue lesions measured by computed tomography (CT) imaging according to modified Response Evaluation Criteria in Solid Tumors (RECIST) or progression on bone scanning according to criteria adapted from the Prostate Cancer Clinical Trials Working Group 2 (PCWG2) guidelines. The clinical end point (overall survival) was defined from the start of ADT treatment or censored at last clinical follow-up.

In addition to plasma, I also accessed metastatic samples obtained at rapid autopsy in the Peter MacCallum Cancer Centre (Melbourne, Australia) CASCADE (Cancer tissue Collection After Death, HREC 15/98, **Table 2.1.2**) program[90]. These patients were recruited close to the end of life, and once the patients had passed away, the rapid autopsy protocol aimed to collect metastatic samples within a few hours of death. Since the current systemic genomic or epigenetic studies such as Cancer Genome Atlas (TCGA) or Encyclopedia of DNA Elements (ENCODE) mainly focused on tumour materials obtained pre-treatment, at diagnostic biopsy or at curative surgery, analysis of the metastatic samples at death could provide important information of end-stage cancer evolution pathways and the causes of treatment resistance.

Table 2.1.1.
Patient plasma sample characteristics

| Trial ID | new ID | sample_type | Baseline(BL) or Progression(PD) | tumour_fraction | PSA level (ng/dL) | targeted genome NGS | high_coverage WGS | targeted methylome | BS-LP-WGS | LP-WGS | targeted methylome on leukocytes | abiraterone (ABI) or enzalutamide (ENZ) | Start of ADT to death (months) | AR-MethSig median methylation ratio | Sample Collection |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| APC21 | 1 | Plasma | BL | 0.891 | 403.5 | Yes | No | Yes | Yes | No | Yes | ABI | 14.2 | low | IRST |
| APC21 | 1 | Plasma | PD | 0.892 | 685 | Yes | No | Yes | Yes | No | Yes | ABI | 14.2 | low | IRST |
| V5322 | 2 | Plasma | BL | 0.8 | x | No | Yes | Yes | Yes | No | Yes | ABI | 59.6 | - | Royal Marsden |
| V5322 | 2 | Plasma | PD | 0.8 | 12.68 | No | Yes | Yes | Yes | No | No | ABI | 59.6 | - | Royal Marsden |
| A3 | 3 | Plasma | BL | 0.74 | 132 | Yes | No | Yes | Yes | No | No | ABI | 55.4 | - | IRST |
| A3 | 3 | Plasma | PD | 0.73 | 215.6 | Yes | No | Yes | Yes | No | Yes | ABI | 55.4 | - | IRST |
| E60 | 4 | Plasma | BL | 0.65 | 67.79 | No | Yes | Yes | Yes | Yes | Yes | ENZ | 95.7 | - | PREMIERE trial |
| A32 | 5 | Plasma | BL | 0.643 | 675.6 | Yes | No | Yes | Yes | No | No | ABI | 23.8 | low | IRST |
| A32 | 5 | Plasma | PD | 0.477 | 611.4 | Yes | No | Yes | Yes | No | No | ABI | 23.8 | low | IRST |
| V4051 | 6 | Plasma | BL | 0.61 | 335 | Yes | No | Yes | Yes | No | No | ABI | 77.2 | - | Royal Marsden |
| E23 | 7 | Plasma | PD | 0.45 | 20 | No | Yes | Yes | Yes | Yes | Yes | ENZ | 40.1 | - | PREMIERE trial |
| E23 | 7 | Plasma | BL | 0.55 | 197.9 | No | Yes | Yes | Yes | Yes | No | ENZ | 40.1 | - | PREMIERE trial |
| A29 | 8 | Plasma | BL | 0.5 | 65.12 | Yes | No | Yes | Yes | No | Yes | ABI | 25.4 | low | IRST |
| A29 | 8 | Plasma | PD | 0.54 | 196.5 | Yes | No | Yes | Yes | No | No | ABI | 25.4 | low | IRST |
| E78 | 9 | Plasma | PD | 0.25 | 5.73 | No | Yes | Yes | Yes | Yes | Yes | ENZ | 47.1 | - | PREMIERE trial |
| E78 | 9 | Plasma | BL | 0.45 | 24.87 | No | Yes | Yes | Yes | Yes | No | ENZ | 47.1 | - | PREMIERE trial |
| O104 | 10 | Plasma | BL | 0.418 | 221.4 | Yes | No | Yes | Yes | No | No | ABI | 39 | - | IRST |
| APC14 | 11 | Plasma | BL | 0.353 | 6.8 | Yes | No | No | Yes | No | No | ABI | 35.6 | - | IRST |
| APC14 | 11 | Plasma | PD | 0.64 | 6.9 | Yes | No | No | Yes | No | No | ABI | 35.6 | - | IRST |
| E36 | 12 | Plasma | PD | 0.18 | 28.49 | No | Yes | Yes | Yes | Yes | Yes | ENZ | 20 | - | PREMIERE trial |
| E36 | 12 | Plasma | BL | 0.325 | 33.02 | No | Yes | Yes | Yes | Yes | No | ENZ | 20 | - | PREMIERE trial |
| A30 | 13 | Plasma | BL | 0.31 | 141.2 | Yes | No | Yes | Yes | No | Yes | ABI | 17.2 | - | IRST |
| A30 | 13 | Plasma | PD | 0.18 | 96.1 | Yes | No | Yes | Yes | No | No | ABI | 17.2 | - | IRST |
| APC27 | 14 | Plasma | BL | 0.301 | 71.86 | Yes | No | No | Yes | No | No | ABI | 50.1 | - | IRST |
| APC27 | 14 | Plasma | PD | 0.227 | 98.43 | Yes | No | No | Yes | No | No | ABI | 50.1 | - | IRST |
| APC19 | 15 | Plasma | BL | 0.227 | 58.53 | Yes | No | No | Yes | No | No | ABI | 20.7 | - | IRST |
| APC19 | 15 | Plasma | PD | 0.013 | 6.89 | Yes | No | No | Yes | No | No | ABI | 20.7 | - | IRST |
| V5074 | 16 | Plasma | BL | 0.21 | 92 | Yes | No | Yes | Yes | No | No | ABI | 58.4 | - | Royal Marsden |
| V5054 | 17 | Plasma | BL | 0.18 | 3150 | Yes | No | Yes | Yes | No | No | ABI | 96.3 | - | Royal Marsden |
| A1 | 18 | Plasma | BL | 0.134 | 3.5 | Yes | No | Yes | Yes | No | No | ABI | 80.9 | - | IRST |
| A1 | 18 | Plasma | PD | 0.57 | 5.89 | Yes | No | Yes | Yes | No | No | ABI | 80.9 | - | IRST |
| APC26 | 19 | Plasma | BL | 0.128 | 126.1 | Yes | No | Yes | Yes | No | No | ABI | 29 | low | IRST |
| APC26 | 19 | Plasma | PD | 0.315 | 488.7 | Yes | No | Yes | Yes | No | No | ABI | 29 | low | IRST |
| A18 | 20 | Plasma | BL | 0.12 | 54.26 | Yes | No | Yes | Yes | No | No | ABI | 60.1 | - | IRST |
| A18 | 20 | Plasma | PD | 0.051 | 83.47 | Yes | No | Yes | Yes | No | No | ABI | 60.1 | - | IRST |
| A20 | 21 | Plasma | BL | 0.096 | 69.89 | Yes | No | No | Yes | No | No | ABI | 35.5 | - | IRST |
| A20 | 21 | Plasma | PD | 0.386 | 7.52 | Yes | No | No | Yes | No | No | ABI | 35.5 | - | IRST |
| A17 | 22 | Plasma | BL | 0.094 | 53.83 | Yes | No | No | Yes | No | No | ABI | 49.1 | - | IRST |
| A17 | 22 | Plasma | PD | 0.218 | 551.9 | Yes | No | No | Yes | No | No | ABI | 49.1 | - | IRST |
| APC9 | 23 | Plasma | BL | 0.046 | 25.98 | Yes | No | No | Yes | No | No | ABI | 104.5 | - | IRST |
| A9 | 24 | Plasma | BL | 0.04 | 277 | Yes | No | Yes | No | No | Yes | ABI | 74 | - | IRST |
| A9 | 24 | Plasma | PD | 0.09 | 50.07 | Yes | No | Yes | No | No | x | ABI | 74 | - | IRST |
| A9 | 24 | Plasma | PD_2 | 0.354 | 684 | Yes | No | Yes | Yes | No | No | ENZ | 74 | - | IRST |
| A9 | 24 | Plasma | PD_3 | 0.774 | 825 | Yes | No | Yes | Yes | No | No | ENZ | 74 | - | IRST |
| APC1 | 26 | Plasma | BL | 0.05 | 4.15 | Yes | No | Yes | Yes | No | No | ABI | 17.5 | - | IRST |
| APC1 | 26 | Plasma | PD | 0.086 | 2.84 | Yes | No | Yes | Yes | No | No | ABI | 17.5 | - | IRST |
| H1 | HV1 | Plasma | HV1_R1 | 0 | x | No | No | Yes | Yes | No | Yes | x | x | - | x |
| H1 | HV1 | Plasma | HV1_R2 | 0 | x | No | No | Yes | Yes | No | Yes | x | x | - | x |
| H2 | HV2 | Plasma | HV2_R1 | 0 | x | No | No | Yes | No | No | Yes | x | x | - | x |
| H2 | HV2 | Plasma | HV2_R2 | 0 | x | No | No | Yes | No | No | Yes | x | x | - | x |

Table 2.1.2.
CASCADE patient and sample characteristics.

| patient | sample_id | metastatic sites | time on ADT (month) | time from ADT to death (month) | AR-MethSig median methylation ratio |
|---|---|---|---|---|---|
| CA27 | M225 | base of bladder | 51.0 | 84.0 | - |
| CA27 | M231 | spine lumbar | 51.0 | 84.0 | - |
| CA27 | M226 | dural base skull | 51.0 | 84.0 | - |
| CA27 | M219 | right adrenal gland | 51.0 | 84.0 | - |
| CA34 | M334 | Liver: right lobe A | 26.5 | 58.5 | low |
| CA34 | M336 | Liver: right lobe C | 26.5 | 58.5 | low |
| CA34 | M337 | Liver: left lobe A | 26.5 | 58.5 | low |
| CA35 | M442 | Left para-aortic lymph node track | 60.0 | 168.0 | - |
| CA35 | M442B | Left para-aortic lymph node track B | 60.0 | 168.0 | - |
| CA36 | M294 | Roof left orbit - brain | 27.0 | 66.0 | - |
| CA36 | M296 | Dura - left orbit | 27.0 | 66.0 | - |
| CA36 | M307 | Liver: right lobe deposit C | 27.0 | 66.0 | - |
| CA43 | M437 | Liver: left lobe nodule 1 | 3.0 | 34.0 | low |

## 2.2 Plasma DNA bisulfite sequencing

### 2.2.1 Plasma DNA extraction

Circulating DNA was extracted from plasma using the Qiagen™ QIAamp Circulating Nucleic Acid kit (or the QIAamp) and quantified using the Quant-iT high-sensitivity Picogreen double-stranded DNA Assay Kit (Invitrogen™) or Qubit Fluorometric Quantification (ThermoFisher™). The QIAamp features efficient DNA purification of circulating nucleic acids from plasma, serum, or urine. The QIAamp protocol comprises of four steps, in principle – lyse, bind, wash and elute. The detailed procedures are as follows:

1) Mix Qiagen™ Proteinase K with plasma samples.
2) Add Buffer ACL and vortex the mixture for 30 seconds.
3) Incubate the mixture under 60 $^{o}$C for 30 minutes.
4) Add Buffer ACB to the lysate in the tube. Close the cap and mix thoroughly by pulse-vortexing for 30 seconds.
5) Place the QIAamp Mini column into the VacConnector onto the QIAvac (the vacuum designed to gently drain the liquid from the column).
6) Switch on the vacuum and have the mixture drawn through the column completely. Switch off the pump and release the pressure to 0atm.
7) Apply 600 µl Buffer ACW1 to the column, and switch on the vacuum again to have the buffer drawn through the column. Switch off the pump and release the pressure to 0atm.

8) Apply 750 µl Buffer ACW2 to the column, and switch on the vacuum again to have the buffer drawn through the column. Switch off the pump and release the pressure to 0atm.
9) Apply 750 µl of ethanol (96%) to the column, and switch on the vacuum again to have ethanol drawn through the column. Switch off the pump and release the pressure to 0atm.
10) Remove the vacuum and discard the connector. Place the column in a clean collection tube. Centrifuge at full speed (20,000g) for 3 minutes.
11) Place the column into a new collection tube and incubate it for 10 minutes.
12) Place the column for a clean new collection tube and apply 120uL dd$H_2O$ onto the centre of the column.
13) Centrifuge at full speed for one minute for elution.

### 2.2.2 DNA sonication

The Covaris™ E220 ultrasonicator with microTUBE plate was used to sonicate germline, solid tumour, or cell line DNA before going into library preparation. The semi-automated machine was suitable for processing multiple samples at the same time. I used the sonication protocol (**see Table 2.2.2.1**) and aimed to achieve DNA fragment size of 180-200 base pairs after sonication.

Table 2.2.2.1.

| Covaris setting for target peak 200 base pairs | | | | |
|---|---|---|---|---|
| Peak incident power | Duty Factor | Cycles per burst | Treatment time | Volume |
| 175 | 10% | 200 | 200 second | 50uL |

### 2.2.3 Bisulfite Conversion

All DNA extracted from either plasma or biopsy samples were treated with ZYMO Gold kit or ZYMO lightning kit (ZYMO™). The samples were mixed with CT conversion reagent, and then placed in the thermal cycler, 98$^{o}$C for 10 minutes, 64$^{o}$C for 2.5 hours, 4$^{o}$C for 30 minutes. LNCaP cell line DNA and germline DNA derived from white blood cells was used to estimate the DNA loss due to bisulfite conversion. Since DNA loses the complimentary strand and becomes single-stranded due to bisulfite conversion, I used the ssDNA Qubit™ fluorometric kit (Thermo Fisher™ Scientific) to estimate the DNA amount before and after the conversion. After bisulfite conversion, the samples were subject to a series of clean-up steps as follows:

1) The bisulfite-converted was added onto the ZYMO-Spin™ IC Column containing 600uL of ZYMO™ M-binding buffer.
2) Centrifuge at full speed (> 10,000 x g) for 30 seconds.
3) 100 µl of M-Wash Buffer then was added to the column. Centrifuge at full speed for 30 seconds.
4) 200 µl of L-Desulphonation Buffer was added to the column and let stand at room temperature (20-30°C) for 15-20 minutes to remove the bisulfite salts.
5) The ZYMO-Spin™ IC Column was subject to 2 times of wash step.
6) The column was placed into a 1.5 ml microcentrifuge tube, 15 µl of dd$H_2O$ was added and centrifuged for 30 seconds at full speed to elute bisulfited converted DNA. I then continued directly to the library preparation.

### 2.2.4 Methylation library preparation

Swift Bioscience™ Accel-NGS Methyl-Seq DNA library kit (Swift Bioscience) was used to perform library construction after bisulfite treatment. The first step of library preparation was to completely denature the DNA and ensure that all DNA fragments became single-stranded. Next ssDNA end repair was used and followed by adaptor ligation. This step recovered most of the fragmented single strand DNA and created the 3'end adaptor tag which allow a complimentary adaptor primer to bind. The extension and ligation steps completed the double-strand DNA molecule with truncated adaptors. The indexing PCR, with optimized PCR cycles depending on the raw DNA inputs, added the full-length adaptor containing a unique sample index and completed the library (**see Table 2.2.4.1**). For raw DNA inputs below 10ng, I used up to 10 PCR cycles, while for inputs between 10-30ng, I used 8 PCR cycles. The completed libraries were subject to Agilent Bioanalyzer for quality control of insert fragment size and library amount and molarity. The library was then ready for whole genome bisulfite sequencing. For library pooling for targeted capture, libraries were then quantified by KAPA library quantification kit (Roche) to accurately estimate the actual number of 'viable' library molecules for sequencing on an Illumina platform.

Table 2.2.4.1.

| PCR cycle numbers for varying input DNA amounts | |
|---|---|
| 1-10 ng plasma DNA | 12 cycles |
| 10-20 ng plasma DNA | 7 cycles |
| 20-50 ng plasma DNA | 5 cycles |
| 20-50 germline DNA | 7 cycles |
| 50-100 germline DNA | 5 cycles |

### 2.2.5 Targeted capture for methylation library

To capture the regions frequently reported to be hypermethylated, I used Roche™ Nimblegen EpiGiant targeted capture kit (Roche) according to the manufacturer's instructions. The capture-based protocol started with blocking the common P5, index, and P7 sequences. The pooled libraries (overall amount: 1200-1500ng) were mixed and the universal blocking oligos and index-specific blocking oligos were added. The mixture of bisulfite converted library, covered with blocking oligo was dried down using vacuum concentrator centrifugal evaporation system to ensure optimal blocking efficiency. The libraries were then ready to be captured. The 75-90 base-pair long capture probes were hybridised with the pooled libraries and, along with hybridisation enhancing buffer the mixture was then transferred to a PCR plate on the thermocycler. The targeted capture step required 64-72 hours (at 47$^{o}$C) to have all the probes binding specifically to the molecules of interest. After the capture, the DNA molecules of our interest were bound to the probes. The probe-captured libraries were quickly transferred and mixed with the Capture Beads on thermocycler with temperature set at 47$^{o}$C. After a 45-minute incubation, the Capture Beads was washed to clear the chemicals used during the capture step. The bead-bound hybridization nucleic acids were then subject to the PCR amplification (13-16 cycles) followed by Beckman Coulter™ Agencourt AMPure cleaning after which the captured libraries were ready to be sequenced.

### 2.2.6 Beads Cleaning

The Beckman Coulter™ Agencourt AMPure PCR Purification System (or the AMPure bead) was used to purify library DNA and remove shorter DNA fragments such as adaptor dimers or truncated adaptors. Depending on the volume ratio of the AMPure beads to PCR product, DNA molecules above certain size bind to the beads. With higher concentration of the AMPure beads smaller fragments binds. During the bead cleanup step, the AMPure beads were added to the library or library mixture. After 5 minutes incubation a magnetic rack was used to separate the beads from the contaminants which were removed by aspiration. Subsequently, 80% ethanol was applied to wash the beads. Low-EDTA buffer (0.1xTE) or $ddH_2O$ was added to elute the purified DNA molecules from the beads.

## 2.3 Pre-processing of methylation NGS data

The sequencing data was returned in a Fastq format, and I performed the bioinformatic procedures with the ultimate aim to extract the level of methylation of all on-target CpG sites. First, I verified the read quality using fastqc, and adapters trimmed using the Trimmomatic v0.36 application. Since DNA methylation primarily occurs on CpG sites, other cytosines residues (Cs) would be converted to thymines (Ts). This phenomenon reduces the complexity of the libraries and makes the mapping more difficult. One solution to address this is to perform read mapping based on three nucleotides (thymine (T), adenosine (A), guanine (G)) with compromised alignment percentage and computational time. I aligned the reads based on three nucleotides to the human genome (hg)19 using the BSMAP v2.90 [91,92]. BSMAP was built based on HASH table seeding algorithm and only searched for locations which mapped perfectly with part of the reads, and this fact largely improved the mapping efficiency. By default, BSMAP does not report the unmapped reads. The output of BSMAP was in SAM file format, and I used picard (picard-tools/2.18.9, http://broadinstitute.github.io/picard) to convert the SAM file into a BAM file which is applicable for downstream data processing.

The next step was to remove duplicated reads. The standard duplication removal algorithm picard which was used to remove duplicates could not handle bisulfite-treated sequences which were non-complimentary. Thus, to remove duplicates, I first used bamtools (bamtools/2.4.0/gnu-4.9.2) to split the top and the bottom reads. I used picard to remove the duplicated reads from the top and bottom separated BAM files separately. Finally, bamtools was used again to merge the top

and the bottom BAM files. In order to keep the quality of the downstream analysis, I only took paired reads into consideration, while un-paired reads were discarded using bamtools.

All the sequencing raw data, including total sequences, mapped reads, percentage of mapped reads (%), bisulfite conversion rate (%), are shown in **Table 2.3.1.** and **Table 2.3.2**. The bisulfite conversion rate (%) was based on each non-CpG cytosine base and calculated by overall thymine counts in non-CpG cytosine sites divided by overall read counts (T/(T+C) *100%).

The size of plasma DNA is between 140-180 base pairs, and thus paired end 100 bp (PE100) sequencing creates a number of overlapping reads. In order to address this all paired, aligned reads were clipped (hard-clipped) using the bamUtil 1.0.13[93] to avoid the potential bias of calling methylation ratio. The CpG methylation ratio of each sample was also based on cytosine reads divided by cytosine plus thymidine reads using BSMAP (see equation).

$$Methylation\ Ratio = \frac{C}{C+T}$$

From all sites included in our predesigned capture panel (Roche Nimblegen SeqCap EpiGiant), only sites with a minimum coverage of 10X were considered for further analysis of CpG methylation status (**Fig. 2.3.1.**). The methylation ratio was computed using the methylKit R package v1.6.2[94].

Fig. 2.3.1.
Box plot showing coverage distribution in target regions by bisulfite high-coverage next-generation sequencing (NGS) in plasma samples

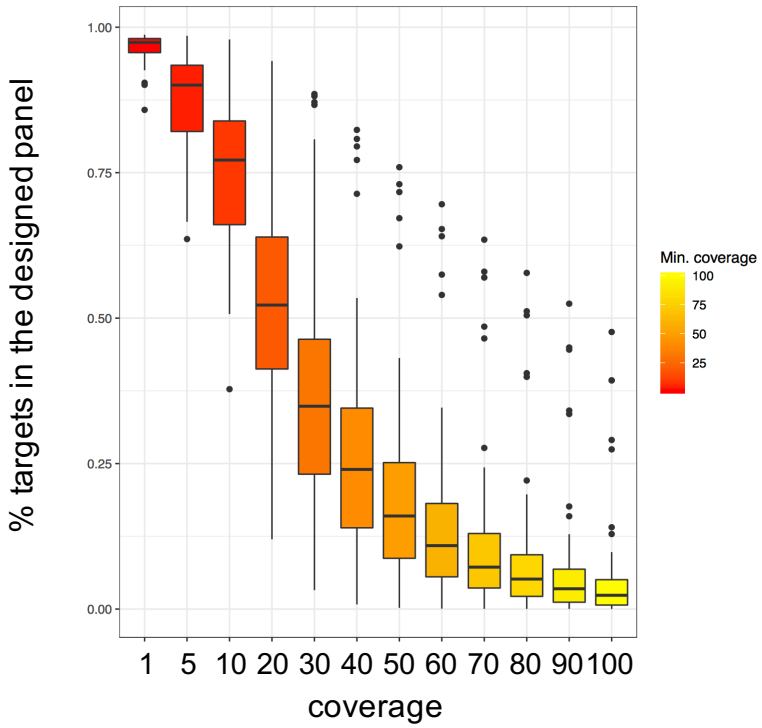


Table 2.3.1.
Targeted methylome sequencing data matrix (total reads, mapped reads, % mapped reads)

| ID | Baseline(BL) or Progression(PD) | sample_type | **targeted methylome** | Total Sequences | Mapped Reads | % Mapped Reads | % bisulfite conversion |
|---|---|---|---|---|---|---|---|
| 1 | BL | Plasma | Yes | 186555220 | 167111560 | 89.58% | 96.6% |
| 1 | PD | Plasma | Yes | 200776658 | 179727102 | 89.52% | 96.5% |
| 2 | BL | Plasma | Yes | 65664842 | 58274518 | 88.75% | 97.0% |
| 2 | PD | Plasma | Yes | 69840940 | 62048086 | 88.84% | 97.0% |
| 3 | BL | Plasma | Yes | 152787676 | 137065257 | 89.71% | 96.7% |
| 3 | PD | Plasma | Yes | 165073482 | 147544256 | 89.38% | 96.5% |
| 4 | BL | Plasma | Yes | 294076770 | 248481849 | 84.50% | 95.4% |
| 5 | BL | Plasma | Yes | 58272348 | 51930705 | 89.12% | 96.7% |
| 5 | PD | Plasma | Yes | 46653168 | 41603630 | 89.18% | 96.6% |
| 6 | BL | Plasma | Yes | 155781620 | 135133214 | 86.75% | 97.1% |
| 7 | PD | Plasma | Yes | 78308978 | 68536718 | 87.52% | 96.5% |
| 7 | BL | Plasma | Yes | 47126584 | 39744701 | 84.34% | 96.2% |
| 8 | BL | Plasma | Yes | 168923880 | 150406167 | 89.04% | 96.6% |
| 8 | PD | Plasma | Yes | 200709408 | 178679515 | 89.02% | 96.5% |
| 9 | PD | Plasma | Yes | 290996960 | 249810225 | 85.85% | 96.5% |
| 9 | BL | Plasma | Yes | 368847482 | 318176786 | 86.26% | 96.6% |
| 10 | BL | Plasma | Yes | 48419738 | 40566381 | 83.78% | 97.1% |
| 11 | BL | Plasma | No | x | x | x | x |
| 11 | PD | Plasma | No | x | x | x | x |
| 12 | PD | Plasma | Yes | 329218080 | 279039378 | 84.76% | 96.4% |
| 12 | BL | Plasma | Yes | 92879856 | 76751623 | 82.64% | 96.2% |
| 13 | BL | Plasma | Yes | 183498796 | 164867778 | 89.85% | 96.5% |
| 13 | PD | Plasma | Yes | 201791470 | 179503148 | 88.95% | 96.6% |
| 14 | BL | Plasma | No | x | x | x | x |
| 14 | PD | Plasma | No | x | x | x | x |
| 15 | BL | Plasma | No | x | x | x | x |
| 15 | PD | Plasma | No | x | x | x | x |
| 16 | BL | Plasma | Yes | 148832250 | 128633440 | 86.43% | 97.0% |
| 17 | BL | Plasma | Yes | 136306032 | 116853097 | 85.73% | 96.9% |
| 18 | BL | Plasma | Yes | 62626728 | 55853347 | 89.18% | 96.7% |
| 18 | PD | Plasma | Yes | 51544194 | 45752062 | 88.76% | 96.8% |
| 19 | BL | Plasma | Yes | 26710136 | 23904934 | 89.50% | 97.2% |
| 19 | PD | Plasma | Yes | 32932662 | 29646068 | 90.02% | 96.9% |
| 20 | BL | Plasma | Yes | 106508740 | 95230613 | 89.41% | 96.8% |
| 20 | PD | Plasma | Yes | 120300158 | 107326545 | 89.22% | 96.7% |
| 21 | BL | Plasma | No | x | x | x | x |
| 21 | PD | Plasma | No | x | x | x | x |
| 22 | BL | Plasma | No | x | x | x | x |
| 22 | PD | Plasma | No | x | x | x | x |
| 23 | BL | Plasma | No | x | x | x | x |
| 24 | BL | Plasma | Yes | 136490762 | 121086940 | 88.71% | 96.6% |
| 24 | PD | Plasma | Yes | 150877238 | 135508492 | 89.81% | 96.6% |
| 24 | PD_2 | Plasma | Yes | 44467698 | 39895075 | 89.72% | 97.1% |
| 24 | PD_3 | Plasma | Yes | 50086012 | 44301319 | 88.45% | 96.1% |
| 25 | BL | Plasma | Yes | 57279412 | 51098610 | 89.21% | 96.5% |
| 25 | PD | Plasma | Yes | 58802244 | 52411556 | 89.13% | 96.4% |
| HV1 | HV1_R1 | Plasma | Yes | 56437928 | 51092703 | 90.53% | 96.9% |
| HV1 | HV1_R2 | Plasma | Yes | 93236514 | 82308864 | 88.28% | 96.6% |
| HV2 | HV2_R1 | Plasma | Yes | 29924470 | 27088855 | 90.52% | 96.7% |
| HV2 | HV2_R2 | Plasma | Yes | 139484410 | 123219652 | 88.34% | 96.9% |

Table 2.3.2.
LP-WGBS data matrix (total reads, mapped reads, % mapped reads)

| ID | Baseline(BL) or Progression(PD) | sample_type | **LP-WGBS** | Total Sequences | Mapped Reads | % Mapped Reads | % bisulfite conversion |
|---|---|---|---|---|---|---|---|
| 1 | BL | Plasma | Yes | 61555740 | 54210537 | 88.1% | 96.6% |
| 1 | PD | Plasma | Yes | 63848530 | 55948225 | 87.6% | 96.5% |
| 2 | BL | Plasma | Yes | 42761332 | 36870167 | 86.2% | 96.0% |
| 2 | PD | Plasma | Yes | 42617996 | 36617189 | 85.9% | 96.4% |
| 3 | BL | Plasma | Yes | 59129422 | 52246290 | 88.4% | 96.7% |
| 3 | PD | Plasma | Yes | 56151334 | 49450718 | 88.1% | 96.5% |
| 4 | BL | Plasma | Yes | 66690658 | 55297446 | 82.9% | 95.4% |
| 5 | BL | Plasma | Yes | 41154970 | 35842986 | 87.1% | 96.7% |
| 5 | PD | Plasma | Yes | 42454336 | 36955379 | 87.0% | 96.6% |
| 6 | BL | Plasma | Yes | 63468228 | 54152133 | 85.3% | 97.1% |
| 7 | PD | Plasma | Yes | 58724038 | 50405549 | 85.8% | 96.5% |
| 7 | BL | Plasma | Yes | 52757540 | 44105950 | 83.6% | 96.4% |
| 8 | BL | Plasma | Yes | 48997884 | 42906582 | 87.6% | 96.6% |
| 8 | PD | Plasma | Yes | 57210482 | 50038902 | 87.5% | 96.5% |
| 9 | PD | Plasma | Yes | 62950726 | 52762980 | 83.8% | 96.5% |
| 9 | BL | Plasma | Yes | 61130412 | 51111528 | 83.6% | 96.6% |
| 10 | BL | Plasma | Yes | 63448740 | 52815910 | 83.2% | 97.1% |
| 11 | BL | Plasma | Yes | 45050378 | 37661639 | 83.6% | 95.8% |
| 11 | PD | Plasma | Yes | 50541554 | 42458690 | 84.0% | 96.0% |
| 12 | PD | Plasma | Yes | 67810208 | 55181787 | 81.4% | 96.5% |
| 12 | BL | Plasma | Yes | 52569972 | 43274790 | 82.3% | 96.4% |
| 13 | BL | Plasma | Yes | 63728198 | 56487135 | 88.6% | 96.5% |
| 13 | PD | Plasma | Yes | 58990260 | 51675897 | 87.6% | 96.4% |
| 14 | BL | Plasma | Yes | 59532904 | 49524710 | 83.2% | 96.5% |
| 14 | PD | Plasma | Yes | 54159938 | 44737438 | 82.6% | 96.8% |
| 15 | BL | Plasma | Yes | 50568866 | 42863081 | 84.8% | 96.0% |
| 15 | PD | Plasma | Yes | 53716688 | 44653610 | 83.1% | 95.9% |
| 16 | BL | Plasma | Yes | 67981380 | 58207701 | 85.6% | 97.0% |
| 17 | BL | Plasma | Yes | 77716154 | 63939783 | 82.3% | 96.9% |
| 18 | BL | Plasma | Yes | 46956924 | 40616653 | 86.5% | 96.7% |
| 18 | PD | Plasma | Yes | 40703206 | 35013233 | 86.0% | 96.7% |
| 19 | BL | Plasma | Yes | 61515246 | 52087141 | 84.7% | 97.3% |
| 19 | PD | Plasma | Yes | 67569626 | 57153038 | 84.6% | 96.9% |
| 20 | BL | Plasma | Yes | 55356132 | 48049486 | 86.8% | 96.8% |
| 20 | PD | Plasma | Yes | 50586228 | 43874943 | 86.7% | 96.7% |
| 21 | BL | Plasma | Yes | 54428970 | 44556958 | 81.9% | 96.0% |
| 21 | PD | Plasma | Yes | 49474000 | 41833481 | 84.6% | 96.0% |
| 22 | BL | Plasma | Yes | 50742732 | 40663027 | 80.1% | 96.5% |
| 22 | PD | Plasma | Yes | 59781268 | 48210928 | 80.6% | 96.5% |
| 23 | BL | Plasma | Yes | 54949210 | 45533981 | 82.9% | 96.5% |
| 24 | BL | Plasma | No | x | x | x | x |
| 24 | PD | Plasma | No | x | x | x | x |
| 24 | PD_2 | Plasma | Yes | 80786148 | 67488386 | 83.5% | 97.1% |
| 24 | PD_3 | Plasma | Yes | 36084166 | 29019487 | 80.4% | 96.1% |
| 25 | BL | Plasma | Yes | 44581236 | 38693632 | 86.8% | 96.5% |
| 25 | PD | Plasma | Yes | 45563992 | 39312802 | 86.3% | 96.4% |
| HV1 | HV1_R1 | Plasma | Yes | 61954404 | 52870463 | 85.3% | 96.9% |
| HV1 | HV1_R2 | Plasma | Yes | 80314646 | 66651587 | 83.0% | 96.6% |
| HV2 | HV2_R1 | Plasma | No | x | x | x | x |
| HV2 | HV2_R2 | Plasma | No | x | x | x | x |

## 2.4 Plasma Methylome Analysis

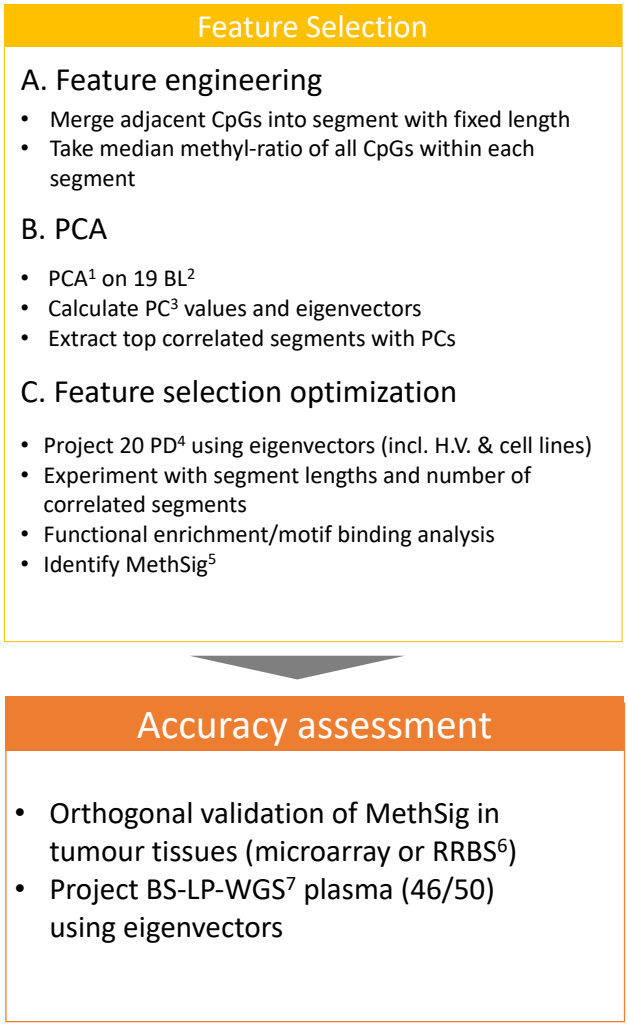
### 2.4.1 Strategies of plasma DNA analysis

Plasma DNA is an admixture of fragmented DNA from different tissues and in cancer patients a majority of plasma DNA arises from tumour cells and leukocytes. Since plasma pan-genome methylation profiles are a combination of the methylation status of cancer cells and other tissues, deconvolution of methylation data can be challenging. Here, I introduce an analysis work flow to identify the plasma methylation signatures related to prostate cancer and to understand the biological processes underlying the methylation signatures (**Fig 2.4.1.1.)**.

First, as the CpG methylation features were often inter-correlated [12] [13] and the main tissue source of plasma DNA from prostate cancer patients came from tumour and leukocytes, I merged the adjacent CpGs into fixed length segments and used the median methylation ratio of all CpGs within the same segment to represent the methylation level of the segment. I then used principal component analysis (PCA) to perform dimensionality reduction in order to understand the variance driving plasma methylome. Each principal component (PC) explained methylation variance in a set of samples. Later, by correlating the PC values with genomically-determined features such as tumour fraction or copy number changes, I aimed to identify methylation features representative of the genomic events. In addition, to fully understand the biological processes of each PC, I extracted segments highly correlated with the PC values and performed functional enrichment analysis to interrogate the common biological pathway driving the

methylation variance. Further, I also conducted motif binding analysis to interrogate the potential epigenetic regulatory factors leading to aberrant methylation phenotypes.

Fig. 2.4.1.1.
Schematic workflow of methylation data analysis.




1. Principal component analysis
2. Baseline plasma samples
3. Principal component
4. Progression plasma samples
5. Methylation Signature
6. Reduced Representative Bisulfite Sequencing
7. Low passage whole genome bisulfite sequencing

### 2.4.2 Principal component analysis of targeted plasma methylome

Methylation segments with methylation ratios available in all baseline samples (*n*=19) and standard deviation values included in the upper two quartiles were subjected to principal component analysis (FactorMineR R package v1.41)[95]. Significant principal components were determined using a permutation test as implemented in the jackstraw R package (v1.2) (https://CRAN.R-project.org/package=jackstraw). The projection of all the samples based on the PCA eigenvectors was based on the methylation ratio of regions used in the initial PCA for the baseline samples. Missing values were imputed based on the PCA method as implemented in the missMDA R package (v1.13)[96].

### 2.4.3 Selection of optimal data inputs for PCA

Adjacent CpG methylation levels are usually highly related, and previously studies have demonstrated high sensitivity of identifying tissue-specific methylation markers using sliding window approaches [12,13,18]. Here I combined adjacent CpG sites into methylation segments of fixed length, and the median methylation ratio across all CpGs within the segment was used to represent the methylation ratio of the segment using methylKit R package v1.6.2[94]. Initially I used 100bps with a sliding window of 50bps and generated > 1.47 million windows across all CpGs in our target panel. I applied principal component analysis (PCA) using the FactoMineR v1.41 package.

To investigate potential bias due to the selection of segmentation length, I optimised the segmentation length parameter. To do so, I tested segments of 10bps, 100bps, 1000bps and 10,000bps with sliding windows of 5bps, 50 bps, 500 bps and 5000 bps, respectively. I found that the smaller the window size, the more data I had to drop when combining plasma samples due to variable inputs and sequencing coverage (**Figure. 2.4.3.1**). I also found that the methylation ratio of 100bps segments with 50 bps sliding window showed high consistency with the methylation ratio estimated at single CpG level (**Figure. 2.4.3.2**). The correlation of PC1 with genomically-determined tumour fraction was >90% regardless of window sizes (**Figure. 2.4.3.3**). Thus, to preserve more detailed methylation information, and to guarantee successful execution in a reasonable amount of time, the setting of 100bps segments with 50 bps sliding window was applied for the rest of our analysis.

Fig 2.4.3.1.
Percentage of data to drop on different window sizes (10bps, 100bps, 1000bps, 10000bps)

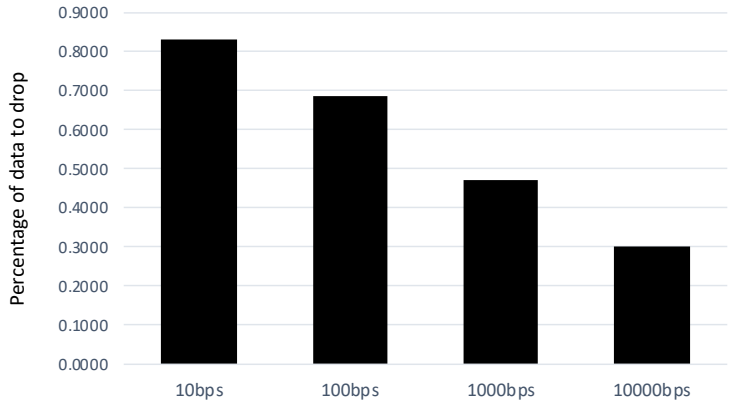


Fig 2.4.3.2.
Distribution of methylation ratio by different segment size (10 bps, 100 bps, 1,000 bps, 10,000 bps)

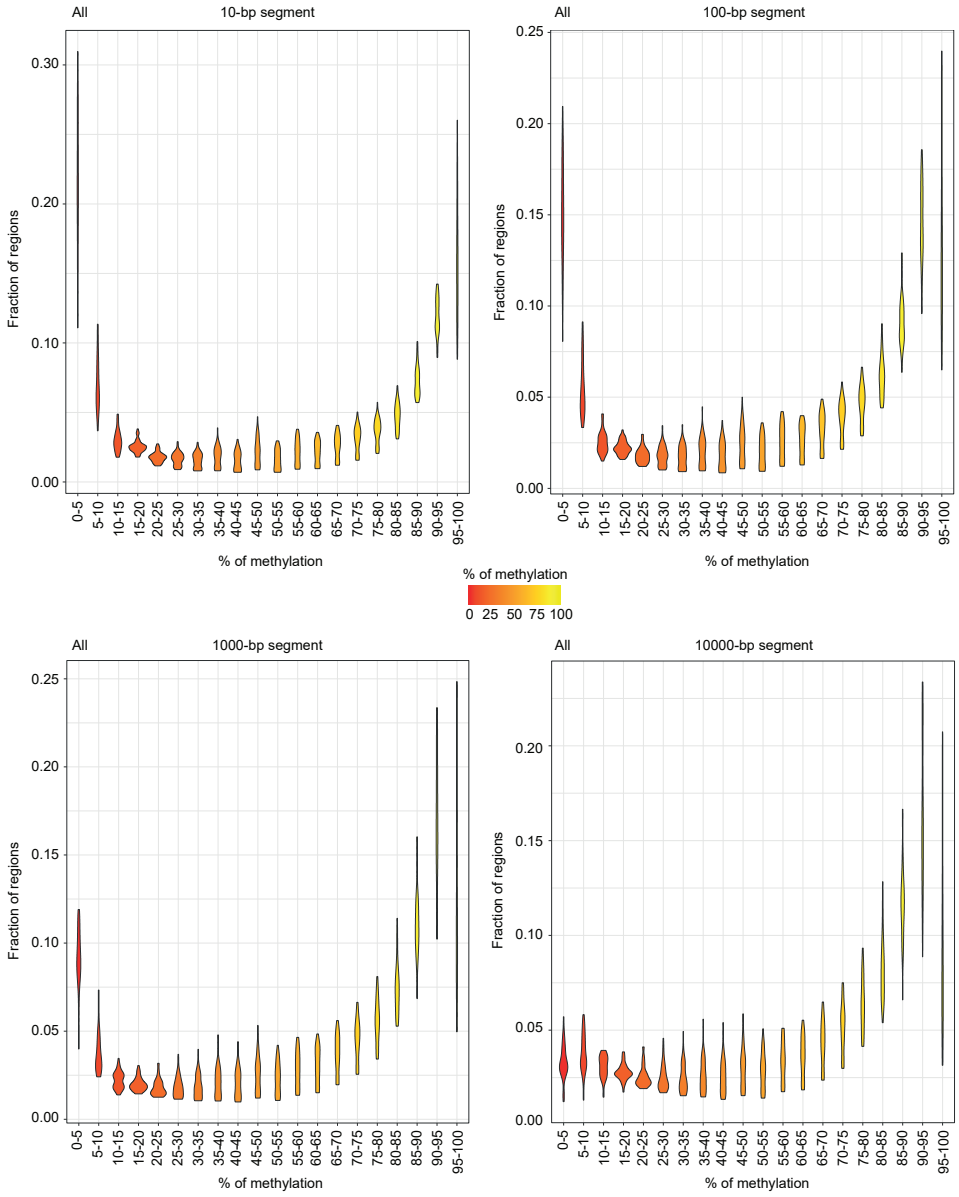
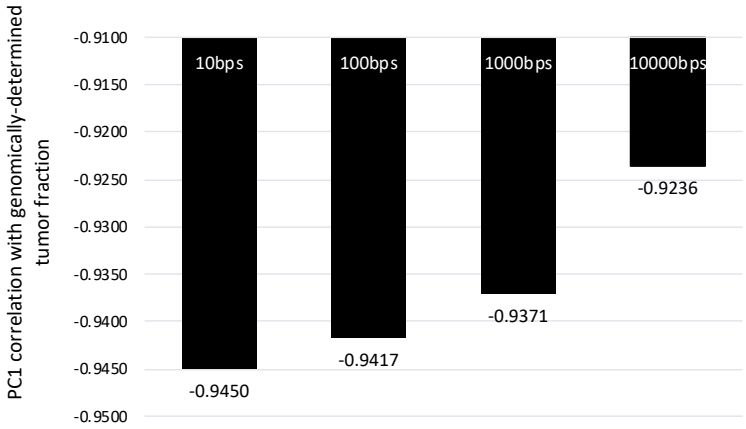


Fig 2.4.3.3.
Correlation of genomically-determined tumor fraction and PC1 values derived from PCA on different window sizes (10bps, 100bps, 1000bps, 10000bps)



### 2.4.4 Methylation Signatures by Principal Component Analysis (PCA)

I applied unscaled PCA using FactoMineR (http://factominer.free.fr) [95]. The PCA model comes with the eigenvector, eigenvalues and correlation matrix comprised of correlation coefficient by each segment. I plotted the distribution of the top-N highly correlated segments based on the correlation matrix returned by PCA, and these segments were highly representative of each eigenvector (e.g., principal component 1, or PC1). To identify the optimal value N of highly correlated segments, I tested multiple N values equal to 10, 100, 1,000, and 10,000 and calculate intra-sample variance, and the correlation between median methylation ratio with genomically-determined tumour fraction (**Figure. 2.4.4.1.**) and the methylation ratio variance of the top-N segments (**Figure. 2.4.4.2.**).

Fig 2.4.4.1.
Correlation of median methylation ratio of selected segments with genomically-determined tumor fraction. Y-axis shows the correlation value and the X-axis denotes the number of top correlated segments

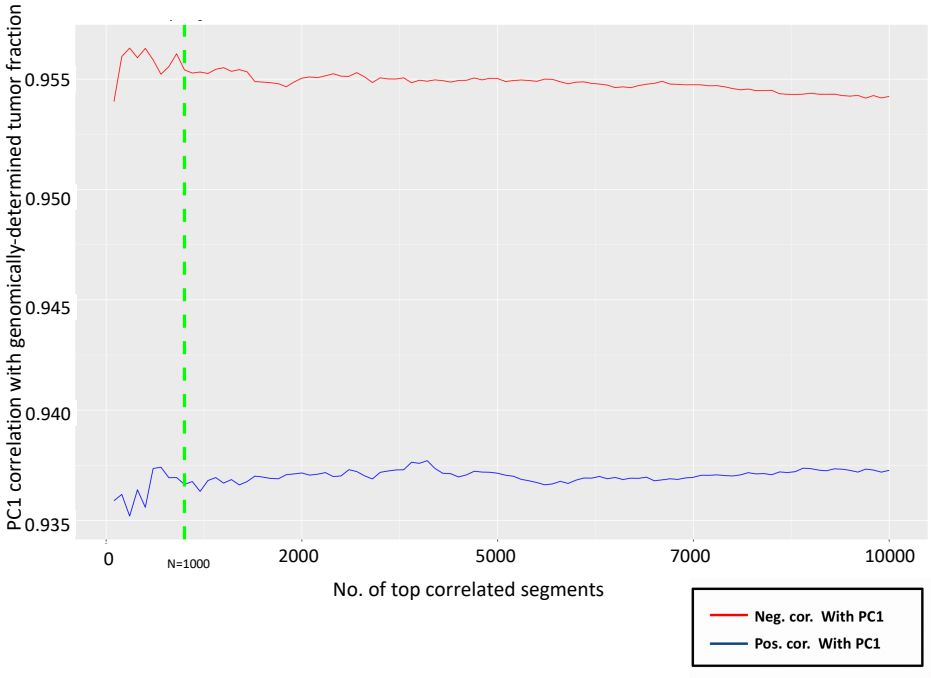
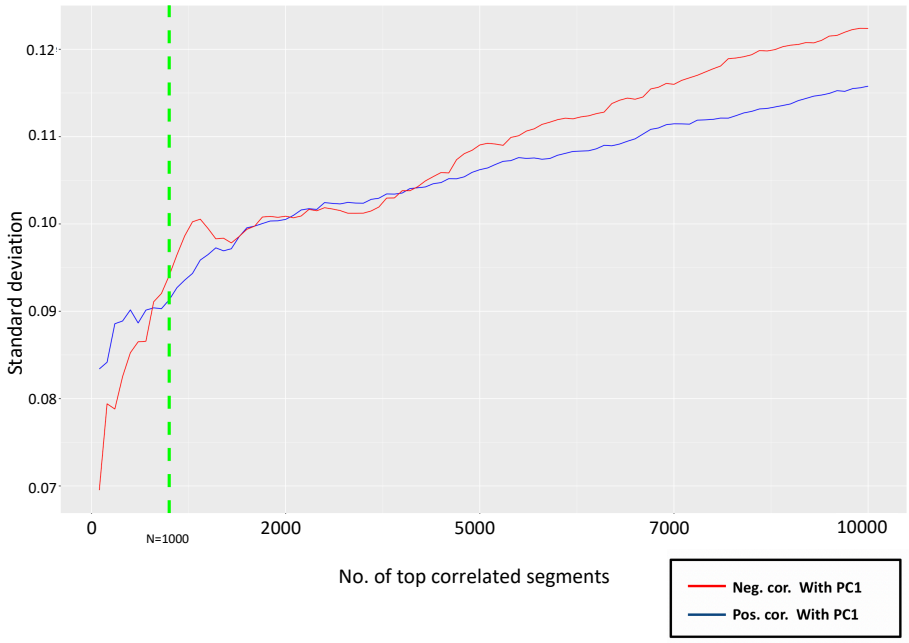


Fig 2.4.4.2.
Standard deviation of methylation ratios of selected segments. Y-axis shows the standard deviation and the X-axis denotes the number of top correlated segments.



### 2.4.5 Gaussian Mixture Model (GMM)

Next, I wondered whether PC1 segments could be divided into prostate specific versus prostate cancer specific. For that reason, methylation ratio of ct-MethSig segments derived from the LNCaP cell line, a normal prostate cell line (PrEC) and healthy volunteer plasma were extracted. To estimate the probability density function (pdf), I applied kernel density estimation (kde), assuming a mixture of two Gaussian distributions consistent with the input dataset of normal prostate epithelium (**Figure. 2.4.5.1**). The Gaussian mixture model (see formula II) applies expectation-maximization (EM) to fit the mixtures of Gaussian distributions by an iterative process [97]. In our experimentation, the model was executed with maximum iterations of 100 times and 'k-means' method for initialization, and I hypothesized that there were two Gaussian distributions, each of them with its own general covariance. The Gaussian mixture model was subject to cross-validation on random split set of regions over 100 times to prove the robustness of the approach (**Figure. 2.4.5.2**). The fitted GMM (number of class = 2) was then used to predict ct-MethSig segments of prostate epithelium (PrEC) [98].

Gaussian mixture model: $g_j(x) = \emptyset_{\theta_j}(x); where\ \theta_j = \left(\mu_i, \sigma_j^2\right)$ (II)

Fig 2.4.5.1.
Distribution of methylation ratio of different tissue types

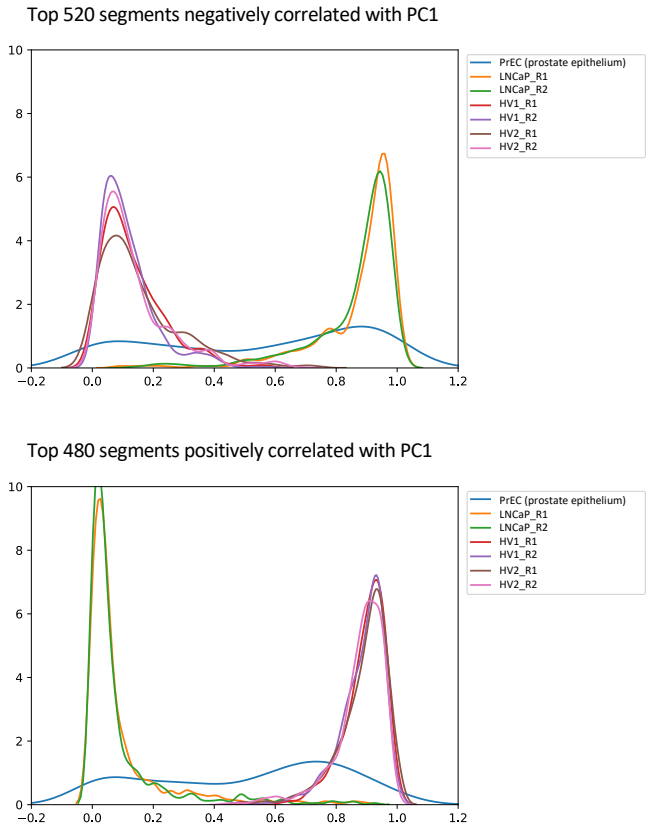
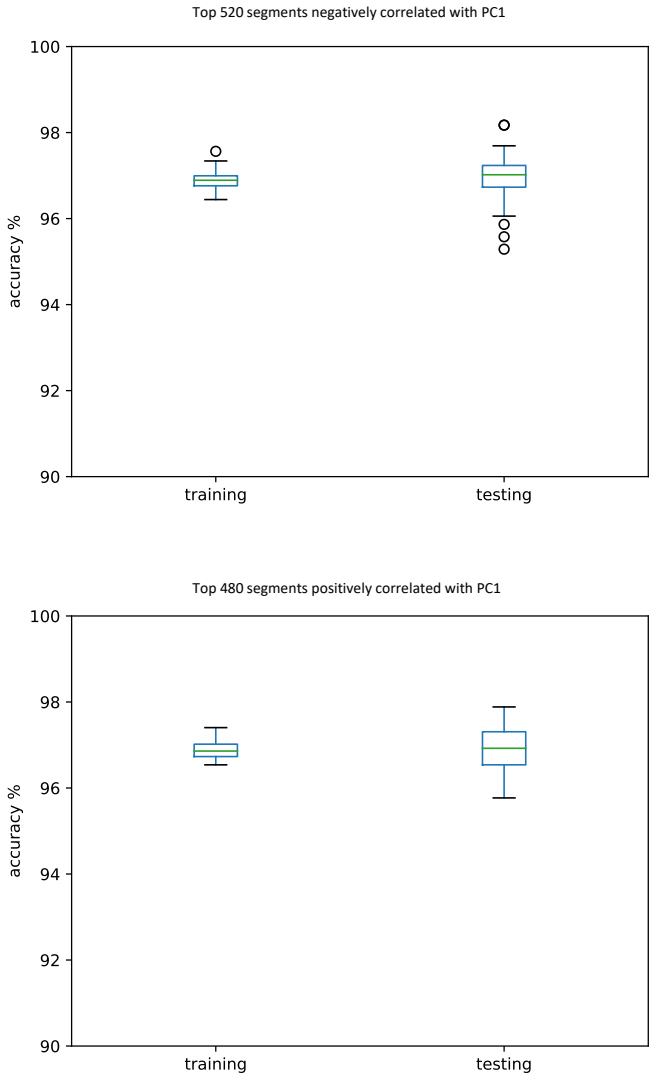


Fig 2.4.5.2.
Performance of Gaussian Mixture Model (k-fold cross-validation, k=100)



### 2.4.6 Classification models

I used the full dataset containing all samples subject to high-coverage targeted methylome to build a classification model to distinguish a plasma sample containing tumour to one without. The plasma methylome derived from metastatic castration-resistant prostate cancer patients and LNCaP cell line methylome were labelled as class 1 (N = 44), while methylome from the white blood cells and that from healthy volunteer plasma samples were labelled as class 0 (N = 19). The dataset was randomly split into training and testing sets (0.75:0.25). I then applied two classic machine learning classification algorithms – random forest classifier (RFC) and Least Absolute Shrinkage and Selection Operator (LASSO) to build a classification model subject to 100-fold cross validation.

RFC model was built using sklearn.ensemble.RandomForestClassifier. The default parameters (eg, max_depth=None, min_samples_leaf=1, min_samples_split=2) were used to control the size of the trees. During the training processes, I experimented on the numbers of trees per forest (N=10, 100, 1000). Each model was applied to predict on testing dataset to measure the accuracy and this allowed us to understand the performance in both training and testing in order to estimate overfitting of the training dataset.

LASSO model was built using sklearn.linear_model.Lasso. The default parameters were kept (eg, max_iter= 1000000), and during the training steps, I experimented on different alpha values which were used to then control the regularisation processes, wherein I penalised the number of features in a model in order to only

keep most important features. For example, the higher the alpha value was, the more the coefficient value of each feature tends to be zero. If the alpha value sets to be zero, LASSO would produce the same coefficients as a linear regression.

## 2.5 Analysis of low-pass whole genome data

### 2.5.1 Low-pass whole genome bisulfite sequencing (LP-WGBS)

Reads from LP-WGBS were processed as methylation high coverage NGS. To calculate PC1 values derived from LP-WGBS, I used the default segmentation length of 100 bps and calculated the methylation ration of each segment (**see 2.4.1 & 2.4.2**). To maximize the available information obtained from our data, I imputed methylation data from higher coverage bisulfite data based regularised iterative PCA algorithm [96] (missMDA R package (v1.13)), and projected on the PCA model as described above. The R package missMDA used a PCA-based model for continuous variables imputation. An imputation process allowed us to obtain results from an incomplete dataset with some missing values such as methylation levels derived from LP-WGBS.

### 2.5.2 Low-pass whole genome sequencing

Low-pass whole genome sequencing (LP-WGS) on untreated plasma DNA was performed with a target of 0.5-1X coverage. For each sample, reads from LP-WGS were aligned to the hg19 using Burrows-Wheeler Aligner MEM algorithm (BWA-MEM) version 0.7.12-r1039. BWA-MEM has been shown to be more efficient and faster than BWA-SW or the BWA-backtrack algorithm. The aligned reads were subject to de-duplication using picard v2.1.0.

## 2.6 Tumour fraction estimation

### 2.6.1 Targeted genomic NGS

Genomically-determined tumour fraction was determined from targeted NGS using CLONET in collaboration with Dr Francesca Demichelis and Dr Alessandro Romanel and were included in three previous publications [73,89,99]. In brief, the CLONET utilised two truncal mono-allelic deletions in prostate cancer – 8q21 and 21q22 to compute the allelic frequency (AF) distribution of the heterozygous single nucleotide polymorphisms (SNPs) across these two regions. The deviation of AF of heterozygous SNPs in plasma sample as compared to that of white blood cells can then be used to infer circulating tumour fraction in a plasma sample.

### 2.6.2 Low-pass WGS with or without bisulfite treatment

I used pan-genome copy number alterations to estimate tumour fraction in collaboration with Dr Mariana Buongermino Perreira who performed the majority of this work. ichorCNA (https://github.com/broadinstitute/ichorCNA) was applied to estimate the tumour fraction using LP-WGS data on bisulfite-treated or non-treated plasma DNA. The human genome was first divided into non-overlapping bins of 1 million base pairs, and, for each sample, the de-duplicated reads were counted per bin using HMM Copy (http://compbio.bccrc.ca/software/hmmcopy/) [100]. The algorithm first removed bins in the centromere regions with a flanking region of 100,000 base pairs. For all the remaining bins read counts were corrected by GC content and mappability issues. The normalised read counts were then fed into the Hidden Markov model

(HMM), which is a probabilistic model assigning each bin into one possible state (hemizygous deletions (HETD, 1 copy), copy neutral (NEUT, 2 copies), copy gain (GAIN, 3 copies), amplification (AMP, 4 copies), and high-level amplification (HLAMP, 5 or more copies). Based on the copy number profile, the model estimated a ploidy and tumour content for every sample. Finally, the algorithm was initiated with ploidy values 2 and 3, and normal fraction, which is 1 minus tumour fraction of 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95. The solution with maximum likelihood among all of these initial combinations was automatically assigned. The CNA status was estimated based on the log ratio (logR) values of each 1Mbp region obtained by the ichorCNA analysis with fixed threshold of 0.5 (GAIN: logR ≥ 0.5, LOSS: logR ≤ 0.5).

### 2.6.3 High coverage whole genome sequencing

To estimate tumor fraction on high coverage data, I worked with Dr Dimitrios Kleftogiannis and Dr Stefano Lise to develop a computational framework that allowed identification of the best combination of ploidy and purity. The computational work was mainly performed by Dr Dimitrios Kleftogiannis. First, we estimated the sequencing coverage for all matched tumor and normal samples in non-overlapping window of 50,000 bp length, using the COV function of DELLY [101]. The COV function of DELLY returned the number of reads per window. The genome coverage per window was normalized (e.g., count of reads from tumor samples divided by count of reads from normal samples) and scaled by the difference of sequencing depth between tumor and normal samples (e.g., tumor

samples were sequenced at ~100X, whereas normal samples at ~30X), and log2 transformed. Note that during the coverage estimation process we excluded the blacklisted bins for hg19 assembly. Using the coverage information mentioned above we assessed manually the ploidy of different genomic segments. To guide our decisions, we considered the coverage levels of sex chromosomes (i.e., all samples are male, thus having one copy of X chromosome in principle) and corrected our estimations by considering all possible copy number segmentations by Sequenza[102].

Next, we generated B-allele frequency (BAF) plots, which informed the allelic intensity of two alleles for all tumor samples, in order to have a better picture of pan-genomic aberrations. We first identified germline single nucleotide variants (SNVs) per patient from the corresponding normal samples. The variant calling procedure was jointly performed using Mutect [103] and Platypus [104]. For the list of germlines SNVs we estimated the BAF values for the corresponding tumor samples. Note, that BAF plots at normal samples would show a flat line at ~0.5 that indicates equal presence of both alleles. However, in tumor samples BAF imbalance was obvious as a result of different ploidy, or copy number aberration; for example, values at 1 or 0 indicated complete absence of one of the two alleles. This information was then integrated with the estimated ploidy from the previous step, to assess tumor content. To smoothen the BAF plots and to refine our variant frequency estimations, we estimated haplotypes with ShapeIT [105] using as input the germline SNVs. We also aggregated all BAF values (e.g., using the median

frequency) of different SNVs that fell into the same bins of 50,000 base pairs for allele A and allele B.

Given the phased BAF plots and the corresponding ploidy estimations, we manually assessed tumor fraction of tumor samples. Since the same BAF values with different ploidy configurations might result in different TC estimations, we considered multiple solutions per chromosome, and selected the solution that better fitted the data across the entire genome. To achieve this, we utilised the input BAF values and applied them on different formulas depending on the most likely ploidy considering loss of heterozygosity, homologous recombination, or all possible configurations for triploid, tetraploid and pentaploid genome. Furthermore, we cross-checked our results with all possible tumor content estimation solutions from Sequenza and selected either the common estimation or the most likely estimation based on both approaches.

### 2.6.4 Plasma methylome measurement

On high-coverage targeted methylation NGS or LP-WGBS data, I calculated PC1 values as described above, and the median of PC1 values extracted from healthy volunteers were set as 0%, while the median of PC1 values derived from LNCaP samples were set as 100% tumour purity. The tumour fraction of all the plasma samples were obtained by interpolation of PC1 projected values.

## 2.7 Analysis of Illumina HumanMethylation450 BeadChip dataset

The microarray processed data were obtained from the Gene Expression Omnibus [106] repository (GSE84043). From the dataset I selected the probes overlapping with MethSig1 segments. The methylation ratio of each segment was obtained considering the median of the β values of the overlapping probes. The tumour fraction estimates by different methods (eg, LUMP [107], pathological reading, and ASCAT [108]) were obtained by the sample information published [38].

## 2.8 Statistical Analyses

### 2.8.1 Methylation ratio difference with Kruskal-Wallis and Dunn's test

The samples were grouped based on tissue of origin and clinical status (white blood cells, plasma healthy volunteer, plasma baseline and plasma progression). Samples were grouped by ct-MethSig and AR-MethSig, and the median methylation ratio of each 100bp segment was estimated in each group of samples. To keep the analysis consistent, I considered only segments present in all samples (340,467 segments). All the selected segments were split in two groups based on the overlap with the promoter region of known genes (263,262 non-promoter segments, 77,205 promoter segments). The promoter region was defined as 1k base-pair upstream and downstream of the transcription start site (TSS). The differences of methylation ratio distribution among each group was calculated using Kruskal-Wallis test (one-way ANOVA on ranks) as implemented in the R v3.4.0 (https://www.R-project.org (2018)) stats package. After I defined the significance of the differences, I assessed the difference of the methylation ratio across each group using the Dunn's test as implemented in FSA R package v0.8.22 (https://github.com/droglenc/FSA).

### 2.8.2 Correlation and association analysis

Correlation analyses of continuous measures were performed using the Pearson correlation method as implemented in the R v3.4.0 stats package. The association

analysis between principal components and CNA of each region was performed by grouping the principal component values of each sample based on the CNA observed for the region (LOSS, NEUTRAL and GAIN). The differences in the principal component values distribution among groups was then assessed using the Kruskal-Wallis test (one-way ANOVA on ranks) as implemented in the R v3.4.0 stats package.

### 2.8.3 Functional enrichment analysis

I performed functional enrichment analysis (chemical and genetic perturbations, MSigDB) was executed using the enrich R package (v0.1) based on all the MSigDB main categories (MSigDB database v6.0) [109] with a significance threshold of 0.05 on Benjamini corrected p values.

### 2.8.4 Motif enrichment analysis

Motif enrichment analysis was used to identify potential transcriptomic regulators of methylation signatures (MethSig). The start of MethSig top 1000 correlated segments were submitted to find the possible motif binding sequences over-represented as compared to the default background set [110]. The pipeline (Pscan-Chip) [110] originally designed for the analysis of chromatin immunoprecipitation followed by next generation sequencing technologies was applied. The program automatically scanned 75 bps preceding and after the first base of each segment included in a methylation signature to look for known transcriptional factor binding motifs obtained from JASPAR 2018, an open-access database of transcription factor (TF) binding profiles. Local enrichment p-value was two-tailed and denoted whether the motif was over-represented in the 150-bp region compared to the genomic regions flanking them. Global enrichment denoted whether the motif binding sequence was over-represented in the region with respect to global background composed of pan-genome putative regulatory regions from various cell lines. I performed the analysis on top highly correlated segments with PC1 or PC3 and other randomly selected regions from our custom,

targeted enrichment panel. The result of ar-MethSig was validated by an orthogonal pipeline [111], and the finding was consistent to original approach as described above.

### 2.8.5 Other statistical analysis

Pearson correlation was used to measure the association between two parameters (principal component values vs genomically determined tumour fraction estimation, or different approaches of tumour fraction estimations). The association between copy number status of each region and principal components was estimated using the Kruskal-Wallis test. Mann-Whitney U test was used to test significant difference of methylation ratio distribution between two groups (*AR* gain versus *AR* non-gain). Hazard ratio in overall survival analysis was calculated using the Mantel-Haenszel method. For all tests, a significance threshold of 0.05 was required unless otherwise specified.

# 3 Chapter 3. Deciphering global plasma DNA methylation variance in metastatic prostate cancer

## Hypotheses

1. It is feasible to profile methylation status of plasma DNA by NGS
2. Plasma methylation status may be driven by tumour purity

## Aims

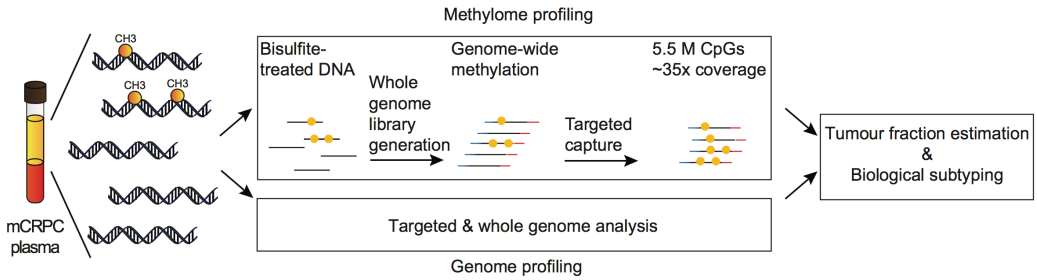
1. To identify plasma tumour methylation using both targeted deep sequencing and low-pass whole genome bisulfite sequencing.
2. To explore methylation difference between tumour and other tissues by statistical procedures, such as PCA.
3. To understand biological consequences underlying key methylation variance.

## 3.1 Interrogating the plasma DNA methylome in metastatic prostate cancer

I concurrently analysed the mCRPC plasma methylome and genome with the aim to accurately quantify tumour fraction and identify distinct biological subtypes (**Fig. 3.1.1.)**. First, plasma DNA was extracted and quantified. In order to identify circulating cell free DNA derived from tumour, I accessed plasma DNA samples that had been subjected to either high-coverage targeted or whole genome NGS. Tumour fractions were determined by quantifying prostate cancer canonical genomic deletions involving 8q21 or 21q22. Separately, copy number status was used for tumour fraction estimation on samples with only LP-WGS data. Determining tumour fractions is crucial for plasma DNA analysis. Here I adapted an algorithm- CLONET [99] to compute tumour fraction by using genomic information at heterozygous single-nucleotide polymorphisms (SNPs) to computationally determine the abundance of deletions involving 8p21 or 21q22, designated as prostate cancer anchor lesions that I had used previously as a proxy for tumour fraction [73,99]. In brief, these 2 regions (i.e. 8p21 & 21q22) were truncally deleted in prostate cancer. In a pure tumour sample where all cells harboured these deletion events I would expect to observe complete loss of heterozygosity across the truncal deletion and thus SNP allelic frequencies of 0 or 100%, while in the normal tissue, the allelic frequency of heterozygous SNPs would be around 50%. Circulating cell-free DNA extracted from plasma and collected from cancer patients is an admixture of tumour and normal tissue DNA

and the allelic frequency of the heterozygous SNPs in the aforementioned regions will be highly correlated with the circulating tumour DNA fraction.

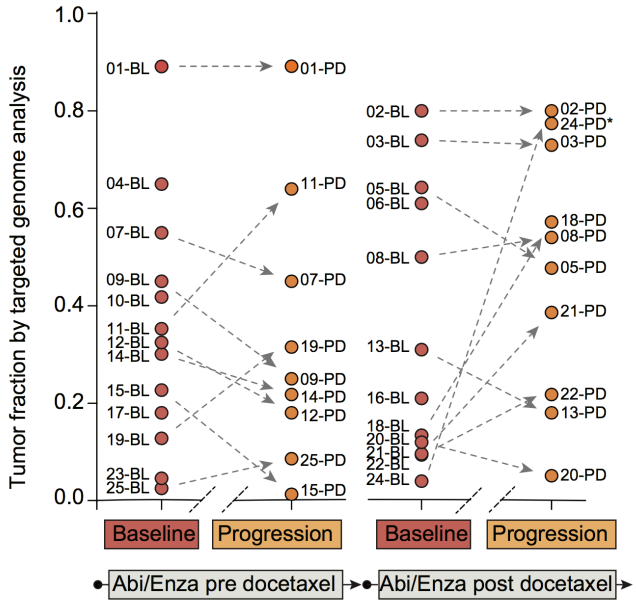
Fig. 3.1.1.
Workflow of integrative analysis of plasma methylome and genome




From the same patient sample, plasma DNA was subject to methylation library generation, and then whole-genome amplified libraries were enriched for 5.5 million CpG sites using the EpiGiant targeted capture kit to characterize the mCRPC methylome. I performed pair-end 100bp deep sequencing on the captured libraries aiming to reach sequencing depth of 35X in average **(Fig. 2.3.1.)**.

All in all, I had plasma samples from 25 mCRPC patients with a wide range of genomically-determined tumour fractions. The patients came from across the mCRPC disease spectrum (docetaxel-naïve or docetaxel-treated) and participated in prospective biomarker protocols collected up to 30 days prior to abiraterone or enzalutamide treatment (baseline). From 19 patients I also had plasma collected at radiographic progression. Additionally, I collected four control samples from two healthy, male volunteers **(Fig. 3.1.2.**, **Table 2.1.1.**). The median and range of tumour fractions in our cohort were: 0.41 (0.04-0.89) and 0.42 (0.09-0.89) for baseline and progression plasma samples, respectively.

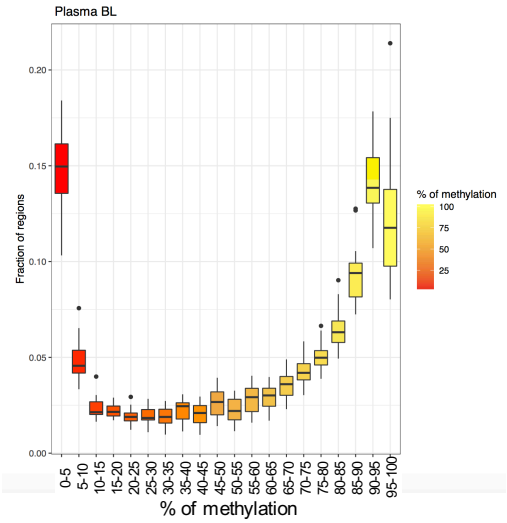
Fig. 3.1.2.
Genomically-determined tumour fraction in baseline and progression samples from pre- and post- chemotherapy patients receiving abiraterone or enzalutamide

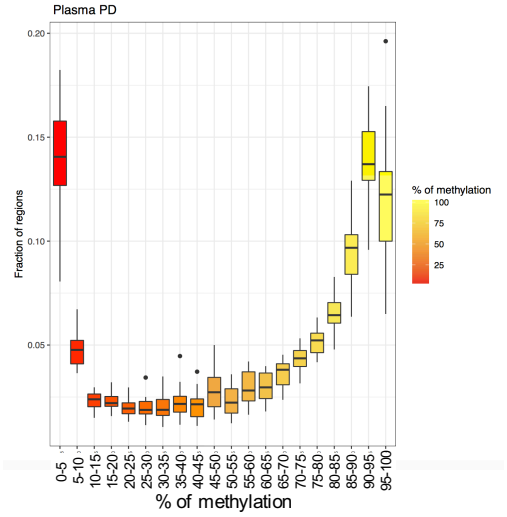



I performed targeted enrichment NGS for 5.5 million pan-genome CpG sites aiming for coverage ≥ 30X on 39 unique plasma samples (19 baseline, 16 progression and 4 healthy volunteer plasma samples from two individuals, **Table 2.1.1.**). I also performed low-pass whole genome bisulfite sequencing (LP-WGBS) on 46 unique plasma samples (24 baseline, 20 progression, 2 healthy volunteer plasma samples from one individual). Additionally, I also conducted targeted bisulfite NGS on DNA from 15 unique white blood cell samples, including 2 samples collected prior to and after treatment with abiraterone from one patient.

Adjacent CpG methylation patterns are usually highly correlated. Therefore, I applied a 100 base-pair sliding window and divided our data into 1.47 million methylation segments. In keeping with prior studies on tissues, the methylation ratio distribution across all methylation segments in plasma and white blood cell samples showed density peaks for hypermethylation (1.0) or hypomethylation (0.0) (**Fig. 3.1.3.**). When separated by annotation category (such as promoter, exon and intron), the distribution of the types of regions captured at 10X coverage was consistent with the regions targeted by the panel (**Fig. 3.1.4.**) [112]. I observed that methylation segments in promoter regions were primarily hypomethylated whilst other categories were primarily hypermethylated (**Fig. 3.1.5.** top panel), and the methylation ratio distributions among all sample types were also significantly different ($P < 10^{-15}$, Kruskal-Wallis test)**.** I then compared the methylation ratio distribution in baseline, progression plasma and healthy volunteer plasma with white blood cell DNA, and I observed significant difference between plasma and white blood cell samples. The difference was more pronounced in cancer patients' plasma samples compared to healthy volunteer ones (respectively, Z scores for promoter regions were −20.3, −19.6 and −15.6 and non-promoter regions: −157.2, −170.1 and −5.9; all $P < 10^{-9}$, Dunn's test, **Fig. 3.1.5.** bottom panel).

Fig. 3.1.3.
Box plot showing methylation ratio distribution for baseline (A) and progression (B) plasma samples and white blood cells (C) presented separately



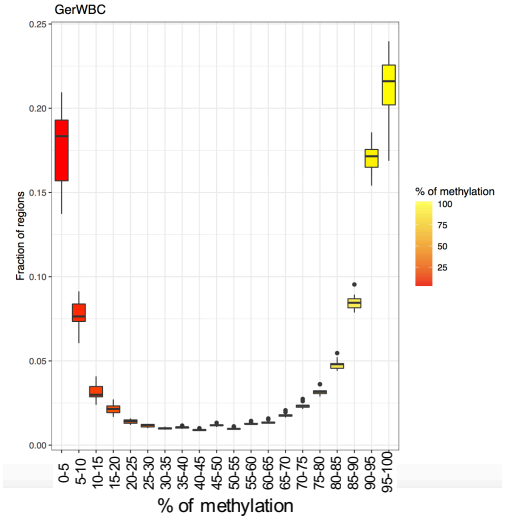



Fig. 3.1.4.
The genomic annotation based on location of methylation segments in the custom targeted panel and in segments covered >10X in 19 baseline samples

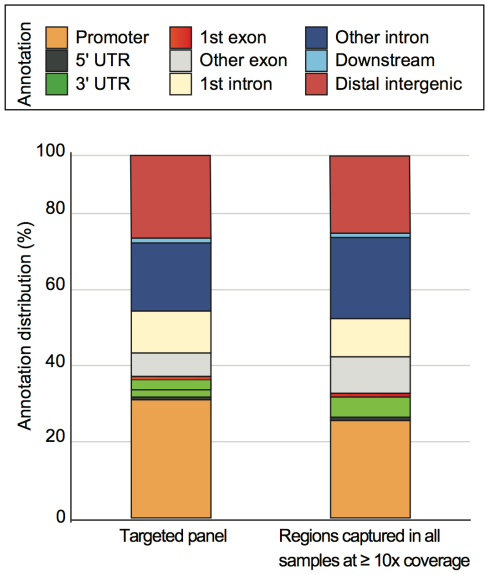
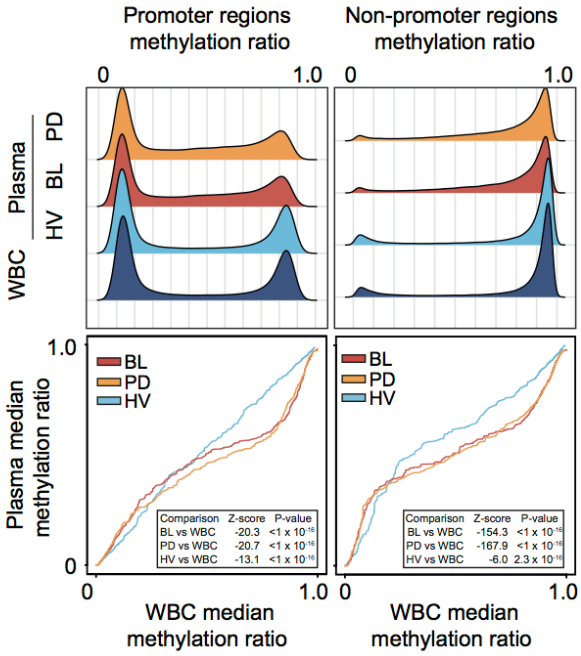


Fig. 3.1.5.
Methylation ratio density (upper panel) and Quantile-Quantile plot (bottom panel) analysis based on the genomic annotation of methylation segments in promoter or other regions.

Data from white blood cells (WBC) or plasma collected at baseline (BL) or progression (PD) from mCRPC patients or from healthy volunteers (HV) are presented separately.



## 3.2 Tumour fraction is the major determinant of global plasma DNA methylation variance

I applied our analytical framework (**Section 2.4.1. and Fig. 2.4.1.1.**) on baseline plasma methylome (*n*=19) to identify methylation features associated with a genomically-determined tumour fraction. To use an unbiased approach to explore the complexity of pan-genome plasma methylation changes, I performed Principal Component Analysis (PCA). I tested different parameters and confirmed the robustness of our finding on progression, healthy volunteer plasma methylome and LNCaP cell line methylome. To expand the applicability of our approach, I extracted segments highly correlated with principal components and tested on LP-WGBS plasma methylome, and external, well-defined tissue data sets using orthogonal approaches such as the Illumina 450k methylation array (**Fig. 2.4.1.1.**).

The first Principal Component (PC1) contributed 42% of the variance (**Fig. 3.2.1.**) and showed a high correlation with genomically-determined tumour fraction (r=–0.96, $P = 1.3 \times 10^{-10}$, Pearson correlation, **Fig. 3.2.2.**). To investigate whether treatment with AR targeting agents affected the association of PC1 with tumour fraction, I used PCA eigenvectors to project the progression samples, healthy volunteer controls ("0" tumour fraction) and the LNCaP prostate cancer cell line ("1" tumour fraction, 3 replicates). After including the projected samples, the correlation of PC1 and genomically-determined tumour fraction remained high (r=–0.94, $P = 1.3 \times 10^{-18}$, **Fig. 3.2.3.**).

Fig. 3.2.1.
Scree plot (top panel) showing principal component analysis (PCA) on 19 baseline samples. Bar-chart shows the variance associated to each Principal Component (PC); the red dotted line indicates cumulative explained variance

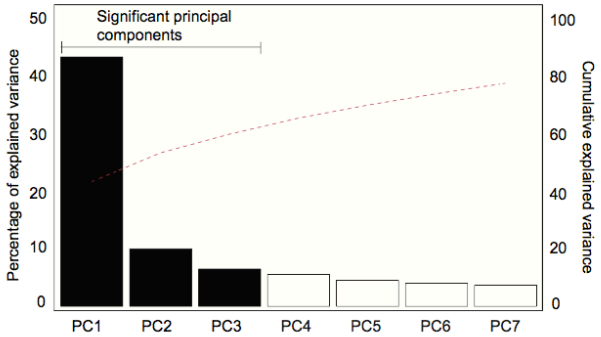



Fig. 3.2.2.
Correlation between PCs and tumour fraction (bottom panel). Size and the colour of each circle show Pearson correlation and background shading denotes P value).

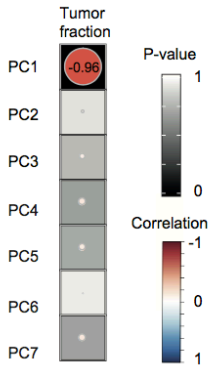
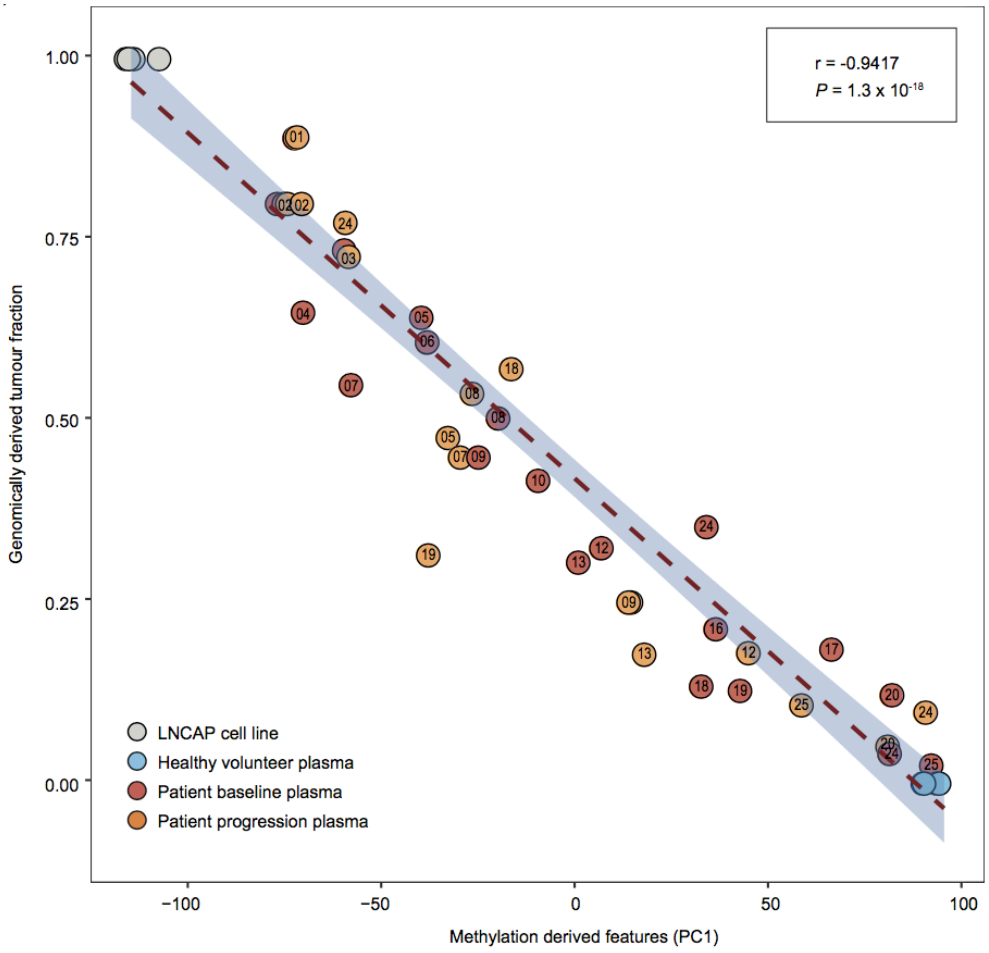


Fig. 3.2.3.
Correlation of genomically determined tumour fraction (y-axis) and principal component 1 (PC1) values (x-axis) from high-coverage targeted methylation sequencing on 19 baseline, 16 progression plasma samples, and control samples (*n*= 4 healthy volunteer plasma samples, LNCaP prostate cancer cell line).



## 3.3 Low-pass whole genome bisulfite sequencing (LP-WGBS)

As part of my methylation analysis protocol I had performed low-pass whole genome bisulfite sequencing (LP-WGBS) which presented a potentially economically efficient and clinically applicable approach for characterising the plasma methylome and extracting methylation signatures. In addition, to have a global overview of methylation changes, LP-WGBS can also profile somatic copy number alterations (SCNAs). It has been suggested that prostate cancer is characterised by complex copy number changes and chromosomal re-arrangements. I hypothesized that methylation features extracted from LP-WGBS can be used to sensitively quantify tumour fraction and complement on copy number aberrations to improve the detection sensitivity.

I first selected an one million base-pair window size to include all the reads within each window and normalized the read number with healthy volunteer plasma samples. Later, I applied the HMMcopy pipeline and visualized the genomic copy number changes across the whole genome to confirm that the copy number alterations detected by LP-WGS and LP-WGBS were similar **(Fig. 3.3.1**). To evaluate the clinical applicability of our findings using LP-WGBS, I then extracted scaled PC1 values from LP-WGBS. Applying Bland-Altman analysis I found a good agreement in the tumour fraction estimates from LP-WGBS compared to that from high-coverage targeted NGS (95% limits of agreement: -0.25 to 0.15, bias: −0.05) introducing the opportunity for scalable and cost-efficient circulating tumour DNA detection and quantitation using LP-WGBS (**Fig.3.3.2.).**

Fig. 3.3.1.
Copy number alteration plots from low-passage whole genome sequencing on plasma DNA with and without bisulfite treatment

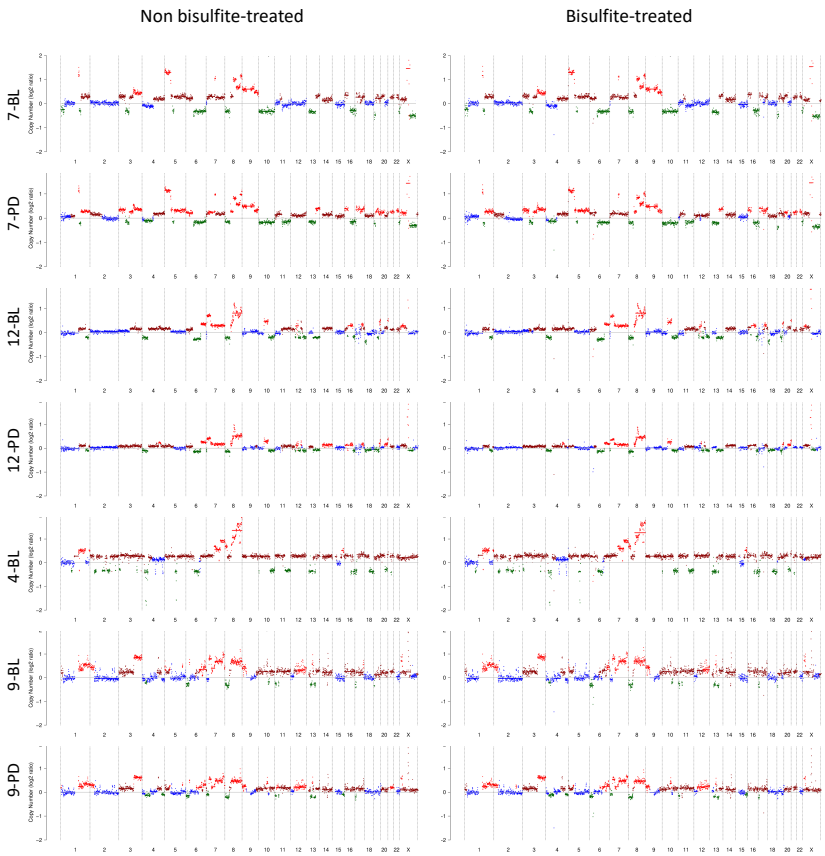
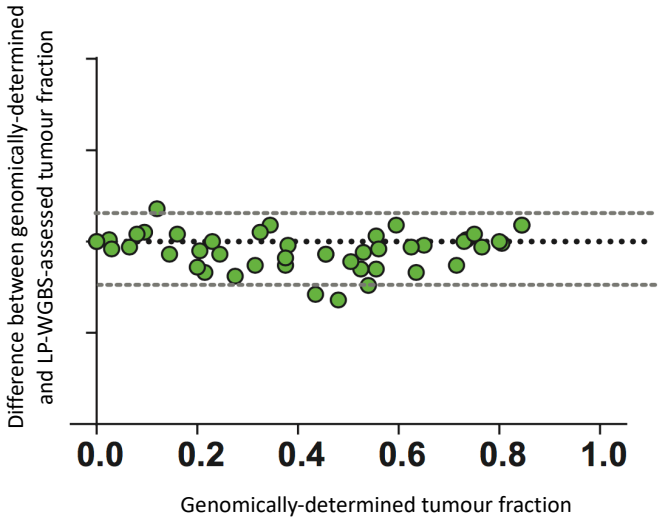


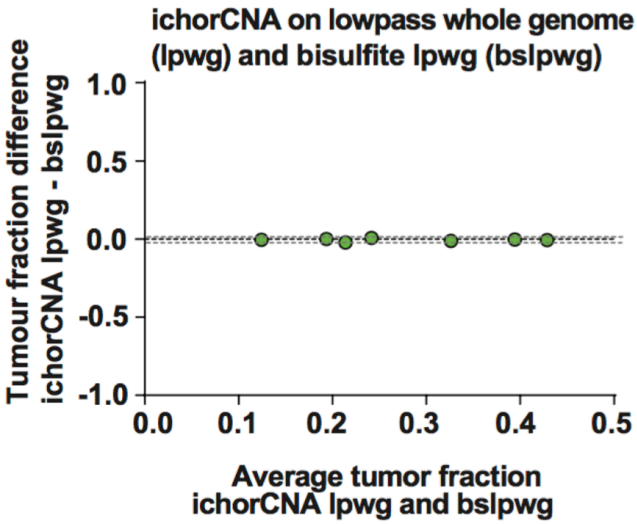
Fig. 3.3.2.
Bland-Altman plot showing agreement between genomically-determined and LP-WGBS-assessed tumour fraction




| | |
|---|---|
| **Bias** | **-0.0480** |
| **SD of bias** | **0.1028** |
| **95% Limits of Agreement** | |
| **From** | **-0.2495** |
| **To** | **0.1534** |

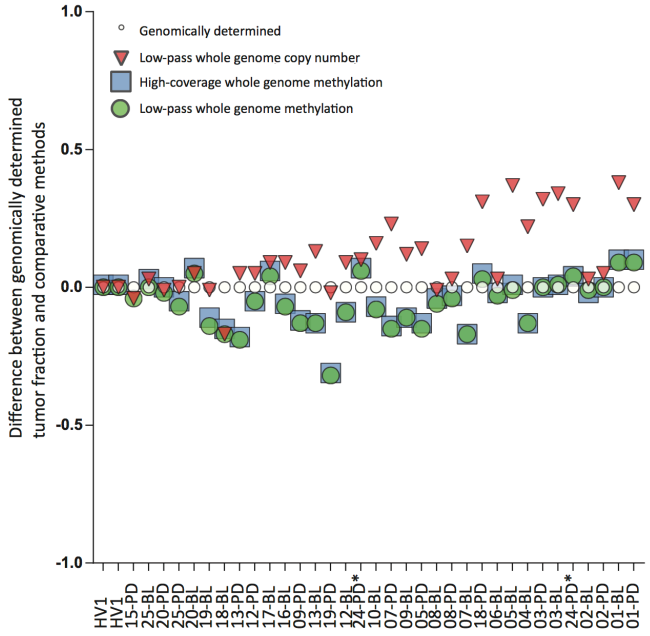
Moreover, I applied a well-validated pipeline (ichorCNA) to estimate the tumour fraction and copy number calls from LP-WGS on untreated and bisulfite-treated DNA from the same plasma samples (**Fig. 3.3.3.**) [75,76]. As a comparison between the copy-number based approach for tumour content estimation (ichorCNA) and the methylation-based measurement, I found that whereas PC1 values showed some over-estimation, ichorCNA tended to under-estimate tumour fraction (**Fig 3.3.4.**). Overall, the result suggested that concurrently extracting methylation values defined in our PC1 with copy number calling from plasma DNA could improve tumour content estimation.

Fig. 3.3.3.
Agreement between tumour fraction estimation by ichorCNA based on LPWG or LP-WGBS




| | |
|---|---|
| Bias | -0.0059 |
| SD of bias | 0.0089 |
| 95% Limits of Agreement | |
| From | -0.0235 |
| To | 0.0116 |

Fig. 3.3.4.
Scatter plot showing the agreement for genomically-determined tumor fraction compared to tumor fraction derived from high-coverage targeted methylome or low-pass whole genome bisulfite sequencing (BS-LP-WGS) or copy number analysis of LP-WGS (ichorCNA). (*multiple progression samples)

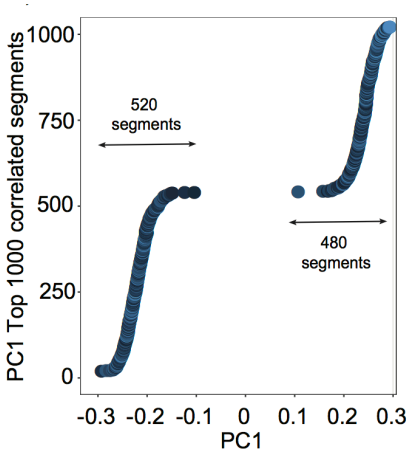


## 3.4 Methylation ratio can serve as a proxy for tumour fraction

To expand the usability of our findings, and test features identified by NGS in datasets with fewer data-points, such as reduced representative bisulfite sequencing (RRBS) or methylation microarray, I hypothesized that the methylation ratios of segments that most strongly correlated to the component features could serve as a proxy of tumour fraction.

I consistently observed a high correlation ($r \geq 0.93$, Pearson correlation, **Fig 2.4.4.1.**) of median methylation ratio with genomically-determined tumour fraction in both negatively and positively correlated group when including 1 to 10,000 segments. Also, the intra-sample variance of methylation ratios in the top correlated segments gradually increased when I included more segments (**Fig 2.4.4.2.**). Therefore, I selected the top 1000 PC1 correlated segments (hereafter referred to as circulating tumour methylation signature, or ct-MethSig) for every plasma sample. In the top 1000 segments, methylation ratios of 520 segments were negatively correlated with PC1, and methylation ratio of the rest 480 segments were positively correlated with PC1 values (**Fig 3.4.1.**). As PC1 values were negatively correlated with tumour fraction, prostate cancer cell line (LNCaP) and high tumour fraction plasma samples presented hyper-methylation features in the 520 negatively correlated segments. On the contrary, prostate cancer cell line and high tumour fraction samples showed hypo-methylation features in the 480 positively correlated segments.

Fig. 3.4.1.
Top 1000 segments (ct-MethSig) with the highest correlation coefficient between PC1 and methylation ratio




Later I confirmed that the median of these methylation ratios showed a high correlation with tumour fraction (520 segments in ct-MethSig hyper-methylated group: r=0.95, $P = 8.4\times10^{-19}$; 480 segments in ct-MethSig hypo-methylated group: r=–0.93, $P = 3\times10^{-16}$, Pearson correlation, **Fig 3.4.2.**). I also tested this finding in published tissue data sets and confirmed a high correlation with tumour purity both in mCRPC [52] (ct-MethSig hyper-methylated group: r=0.92, $P < 1.5\times10^{-6}$; ct-MethSig hypo-methylated group: r=–0.74, $P < 1.4\times10^{-3}$, Pearson correlation, **Fig 3.4.3.**), and hormone-sensitive prostate cancer (HSPC) [38] (ct-MethSig hyper-methylator group: r=0.907, $P < 10^{-60}$; ct-MethSig hypo-methylator group: r=–0.61, $P < 10^{-17}$, Pearson correlation) (**Fig 3.4.4.**). Intriguingly, the methylation-based tumour fraction estimation (LUMP) showed the highest concordance with ct-MethSig median methylation ratio measurement among all different tumour

content estimation pipelines [38] (**Fig 3.4.5.**). Additionally, ct-MethSig did not include genes whose methylation status has been previously reported as diagnostic of prostate cancer as the segments overlapping with these genes were not as strongly correlated with principal component 1 value as ct-MethSig (**Fig. 3.4.6**).

Fig. 3.4.2.
ct-MethSig methylation ratio distribution by patient plasma sample split by negatively correlated segments (hyper-methylator group) and positively correlated segments (hypo-methylator group).

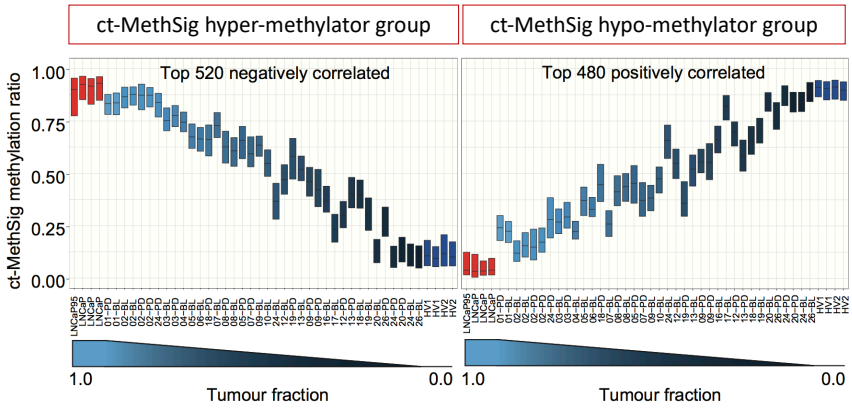



Fig. 3.4.3.
ct-MethSig segment methylation ratio split by hyper-methylator and hypo-methylator groups derived from mCRPC tissues lined by tumour fraction

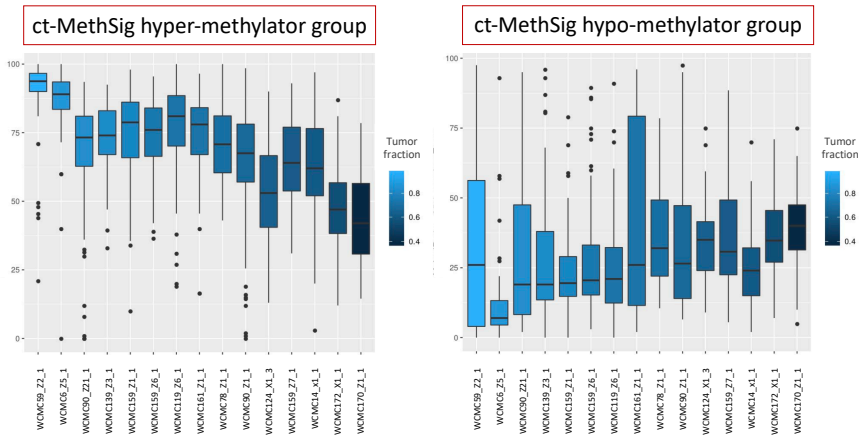


Fig. 3.4.4.
Correlation between HSPC tissue tumour fraction estimation by ct-MethSig and molecularly-defined tumour fraction

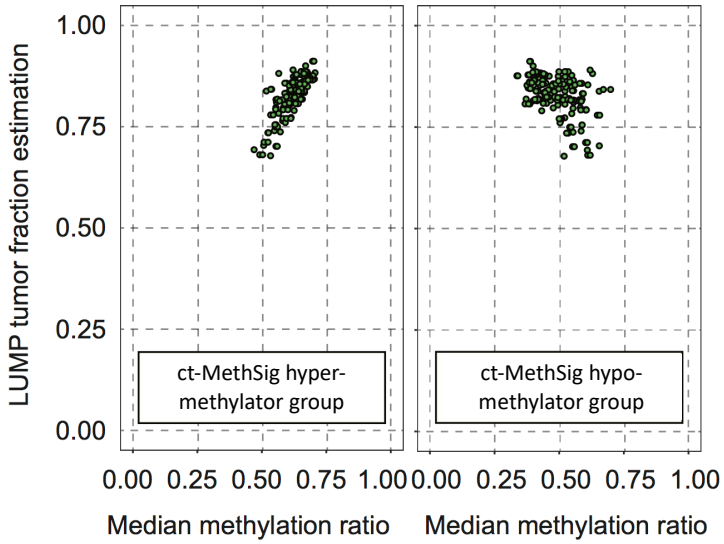


Fig. 3.4.5.
Correlation of median methylation ratio of MethSig1 segments from hormone-sensitive prostate cancers using pathologic (A), and Qpure (B), ASCAT (C) estimates

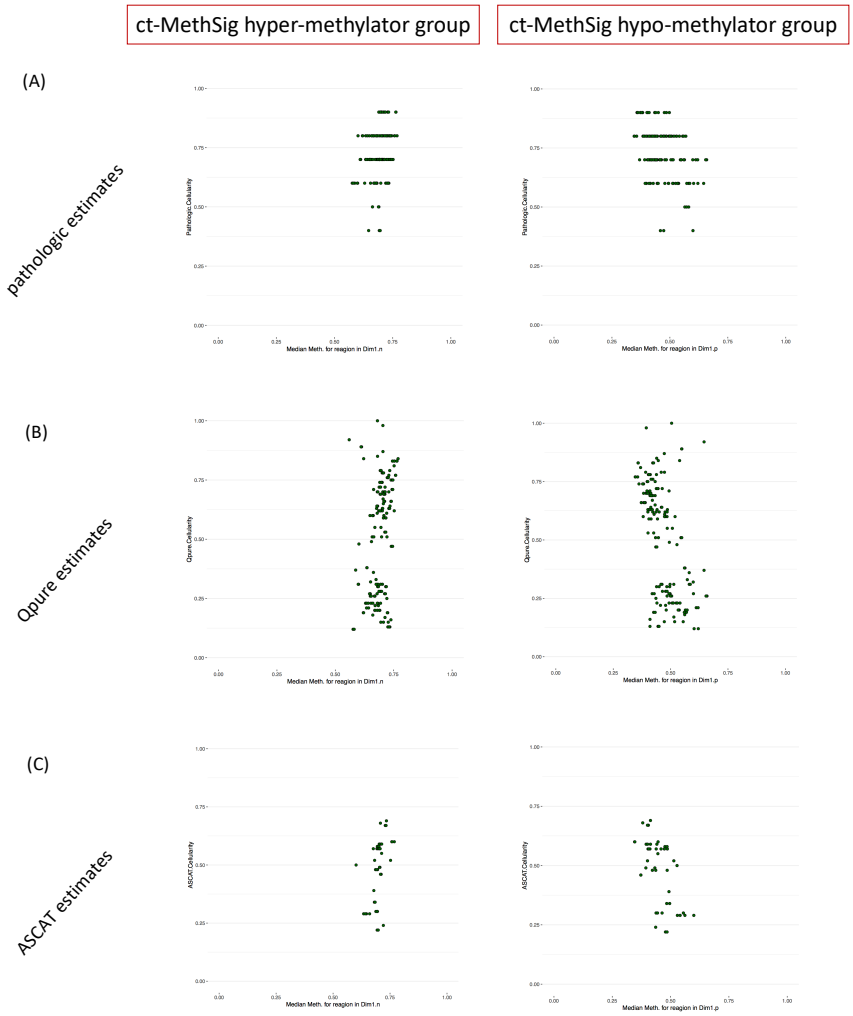
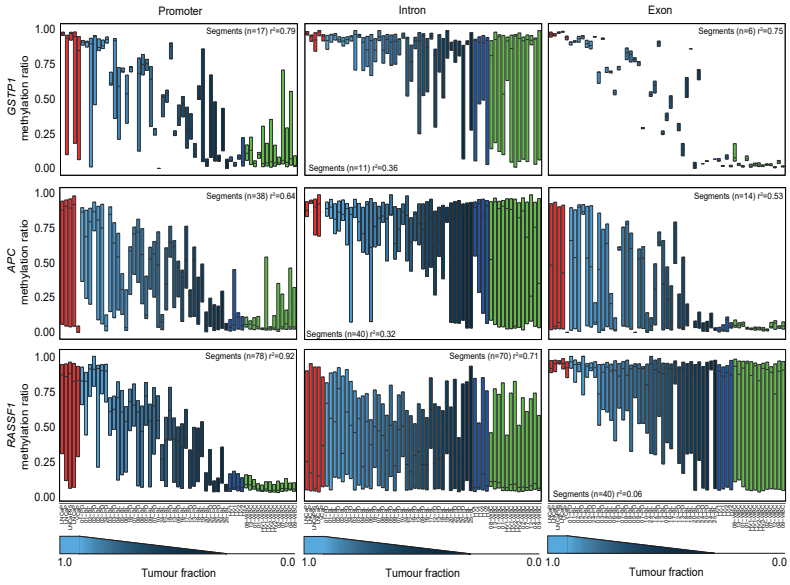


Fig. 3.4.6.
Methylation ratios of GSTP1, APC, and RASSF1A across different tissue types—healthy volunteer plasma, white blood cells, CRPC plasma samples, LNCaP cell line.




In summary, methylation ratios of ct-MethSig can be used as a proxy of tumour fraction, and ct-MethSig hyper-methylator group tended to give a better tumour fraction estimation than ct-MethSig hypo-methylator group across different datasets.

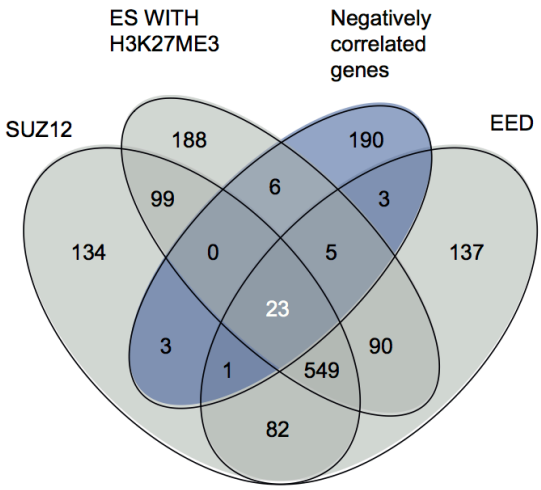
## 3.5 Functional enrichment identifies hypermethylation of polycomb repressor complex 2 targets in circulating prostate cancer DNA

To understand the biological processes underlying PC1, I performed functional enrichment analysis (chemical and genetics perturbations, MSigDB) on genes overlapping with ct-MethSig segments. I observed significant enrichment (adjusted $P < 10^{-4}$) for targets of the polycomb repressor complex 2 [113] (PRC2 related category in the Molecular Signature Database or MSigDB , **Fig. 3.5.1.**) that was of particular interest as a previous study on advanced prostate cancer tissue showed that cancer was distinguished from non-cancer epithelium by down-regulation of genes that are repressed by PRC2 [114]. I noted that this PRC2 enrichment only involved negatively correlated methylation segments that represented an increase in methylation ratios with increasing tumour fraction. The 520 negatively-correlated segments included 231 genes. Of these, 41 were collectively targets of EED (Embryonic Ectoderm Development) and SUZ12 (suppressor of zesta 12) or H3K27ME3 (tri-methylation of lysine 27 on histone H3 protein subunit, **Fig. 3.5.2.**). This discovery of hypermethylation in promoters upstream of these genes provides a biological explanation for their down-regulation and potentially introduces a strategy for extending this biological difference to a liquid biopsy clinical application [52 114].

Fig. 3.5.1.
Functional enrichment analysis of genes (*n* = 231) in ct-MethSig segments. The p-value was corrected for multiple statistical testing (Benjamini-Hochberg).

| | Gene set id | ct-MethSig enriched gene set | P-value adjusted | Genes input/ background |
|---|---|---|---|---|
| Input: negatively correlated genes | M10731 | BENPORATH_ES_WITH_H3K27ME3 | $1.43 \times 10^{-07}$ | 34/1118 |
| | M7617 | BENPORATH_EED_TARGETS | $4.49 \times 10^{-07}$ | 32/1062 |
| | M8448 | BENPORATH_PRC2_TARGETS | $1.03 \times 10^{-05}$ | 23/652 |
| | M16955 | LIVER_CANCER_WITH_H3K27ME3 | $4.44 \times 10^{-05}$ | 13/228 |
| | M9898 | BENPORATH_SUZ12_TARGETS | $1.61 \times 10^{-04}$ | 27/1038 |
| Input: positively correlated genes | M6441 | HCMV_INFECTION_18HR | $1.31 \times 10^{-02}$ | 8/204 |
| | M14437 | AML_CLUSTER_5 | $1.31 \times 10^{-02}$ | 4/40 |
| | M14791 | COLORECTAL_ADENOMA | $4.23 \times 10^{-02}$ | 9/324 |
| | M1949 | NPC_HCP_WITH_H3K4ME2 | $4.23 \times 10^{-02}$ | 10/393 |

Fig. 3.5.2.
Venn diagram of showing the overlap of negatively (dark blue) correlated genes in ct-MethSig segments with targets of EED, SUZ12, and ES (Embryonic Stem cells) with H3K27ME3 marks. The numbers highlighted in white bold denote the number of genes in the ct-MethSig negatively correlated group



## 3.6 Circulating tumour methylation signature comprises segments specific to either normal or malignant prostate epithelium

I posited that ct-MethSig included components that were specific to either prostate malignant or non-malignant epithelium. I plotted the kernel density estimation of the ct-MethSig methylation ratios in whole genome bisulfite sequencing data derived from the non-malignant prostate epithelium cell line (PrEC) [115] and I observed that there was a bimodal distribution (**Fig. 3.6.1.**). I therefore adapted Gaussian mixture model on methylation ratios of ct-MethSig segments from the prostate cancer cell line LNCaP and our two healthy volunteer plasma samples and then I used the fitted Gaussian distribution on normal prostate epithelium (PrEC). In PrEC I identified segments whose methylation ratio distribution aligned with either LNCaP or healthy volunteer plasma. I concluded that the former segments with methylation ratios in normal prostate epithelium similar to LNCaP were prostate epithelium-specific, while the segments with methylation ratios similar to healthy volunteer plasma were prostate cancer-specific (**Fig. 3.6.1.**). I then confirmed these findings by showing that CRPC metastases (bone, bladder, liver and lymph nodes, described further in Supplementary Table S4) included segments attributed to both normal and cancerous prostate epithelium whilst normal prostate (54 year-old male donor, ENCODE donor ID: ENCDO451RUA) included only segments attributable to normal prostate epithelium. As a result, I could therefore split ct-MethSig into two components, circulating cancer-specific and normal prostate-specific signatures. Finally, I used methylation microarray data from 553 prostate cancers from TCGA

and 12 CRPC adenocarcinoma from Beltran et al. [52] to show that the distribution of ctMethSig segments in localized prostate cancer and CRPC tissue includes both cancer and normal components (F**ig. 3.6.2.**).

Fig. 3.6.1.
Circulating tumor fraction methylation signature comprises segments specific to either normal or malignant prostate epithelium.

Left panel: Methylation ratios of ct-MethSig negatively (N=520) and positively(N=480) correlated group from LNCaP(N=4), healthy volunteer (H.V., N=4), and normal prostate epithelium (PrEC)

Right panel: ct-MethSig negatively and positively group can be split into prostate cancer specific segments and prostate epithelium specific.

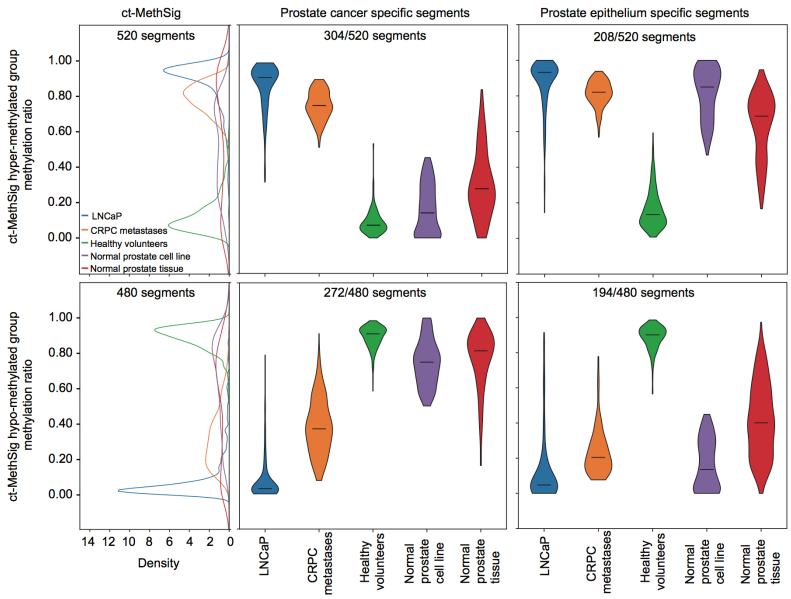
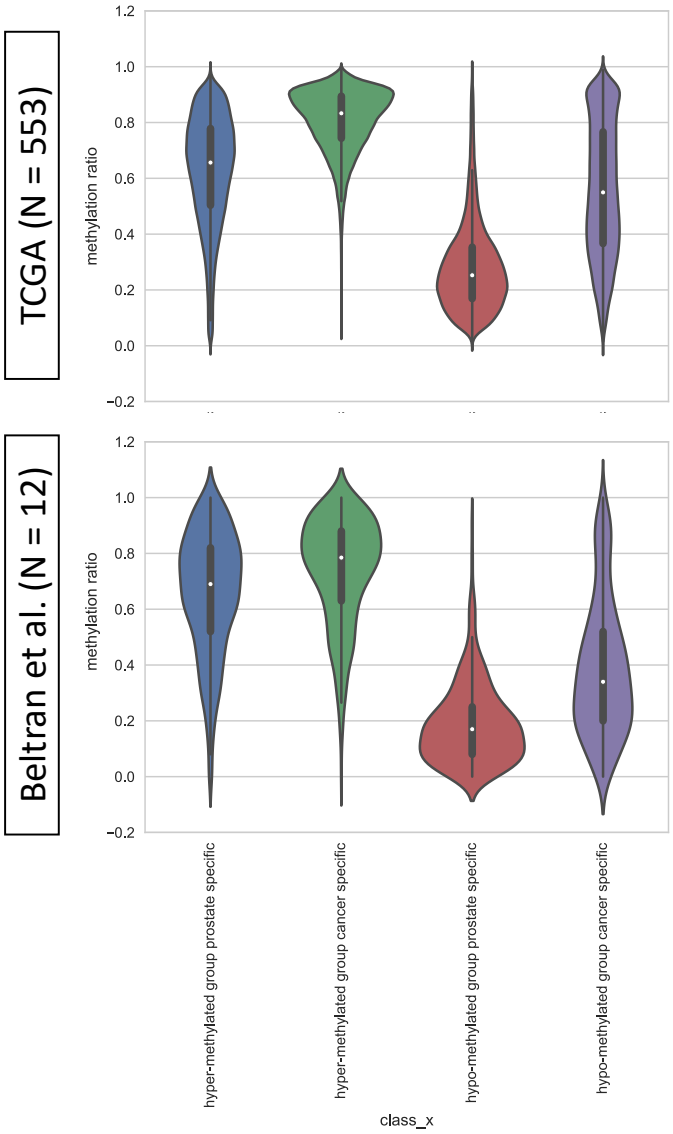


Fig. 3.6.2.
Methylation ratio distribution of circulating normal prostate specific or prostate cancer specific component in localized prostate cancer from TCGA



## 3.7 Principal component 2 was driven by a single patient and can be associated with tumour with distinct genomic aberrations.

The second principal component (principal component 2 or PC2) represented 10.1% of global plasma methylation variance. When I looked into the top 1000 segments which were highly correlated with PC2, I observed that only patient 02 (clinical trial ID: V5322, see **Table 2.1.1.**) showed relative hypo-methylation patterns (**Fig. 3.7.1.**). PCA contribution matrix confirmed that patient 02 contributed significantly to PC2 (**Fig. 3.7.2.**). In general, each sample may only contribute < 10% of each principal component; however, this patient contributed >60% of the principal component two.

Figure 3.7.1.
Methylation ratio across PC2 top 1000 highly correlated segments

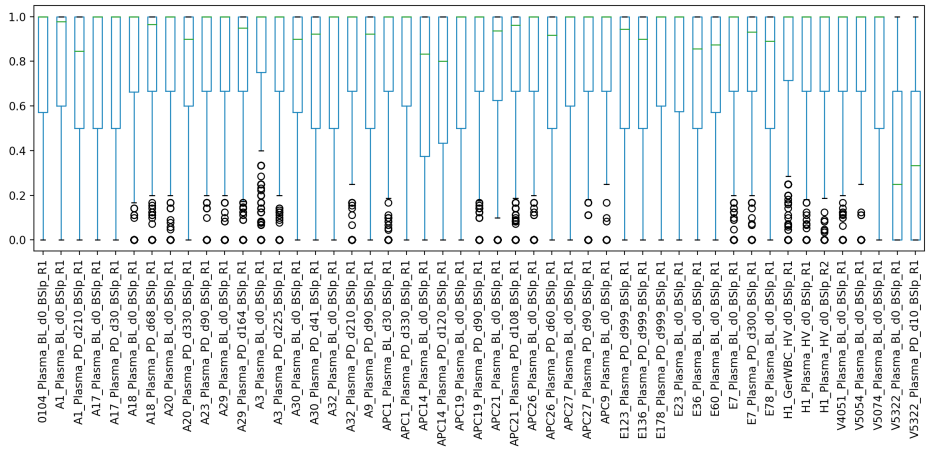
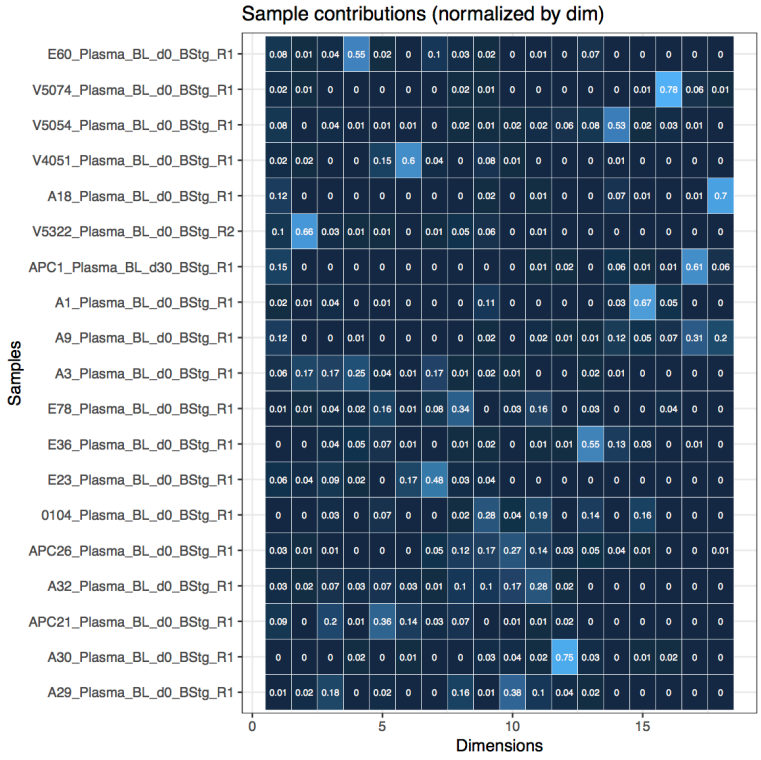
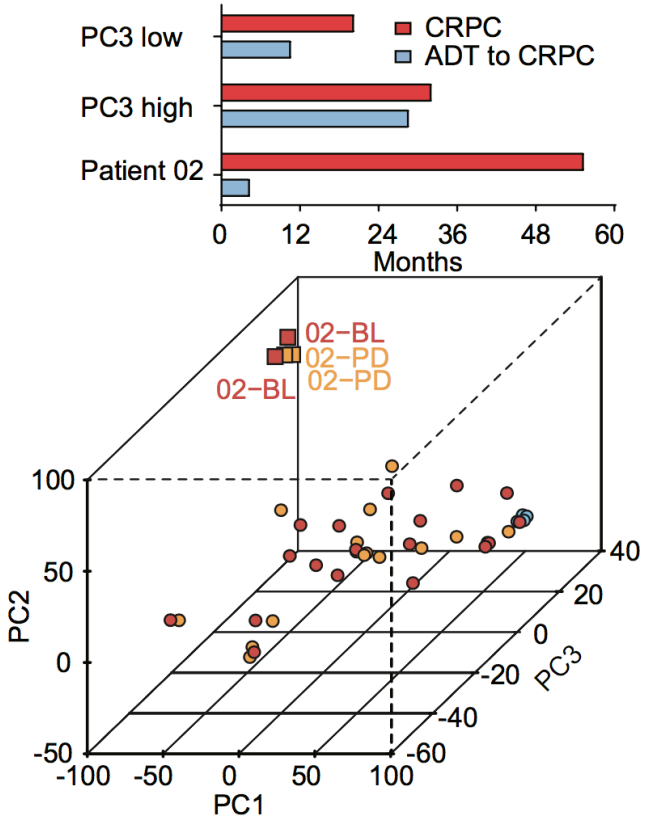


Figure 3.7.2.
Principal component analysis contribution matrix by plasma samples




On further investigation, this single patient was found to be a clinical outlier: the patient had a very short interval from androgen deprivation therapy (ADT) to mCRPC (<4 months) and no response to standard treatment with docetaxel, cabazitaxel or enzalutamide administered in this sequence. The presenting prostate specific antigen (PSA) was low (8.2 ng/dl) despite high volume de novo metastatic disease and there was strongly-positive AR and PSA staining on liver biopsy. Interestingly, the patient had an exceptional response to carboplatin lasting longer than 30 months (**Fig. 3.7.3.**). However, no genomic aberrations in DNA repair genes were detected on whole-genome sequencing of his plasma DNA.

Figure 3.7.3.
Visualisation of three orthogonal principal components (PC1, PC2, PC3). The inset indicates the difference of clinical outcomes of PC3 high and PC3 low groups. For comparison the PC2 outlier (patient 02) is included. Blue bar denotes median time from start of ADT to development of CRPC and the red bar denotes median time from development of CRPC to death.




Further, I analysed the whole genome sequencing data of the patient samples (both baseline and progression) and found that they had high beta allelic imbalance resembling chromotripsis (**Supplementary Fig. 3.7.4.**). When I plotted the beta allelic frequency map across the whole genome, PC2 top correlated segments were more sparse across the whole genome as compared with PC1 top correlated segments. Some PC2 top correlated segments were located in the regions (e.g., chr6, chr8 and chr17) with high allelic imbalance which may be due to deletion or inversion events. Moreover, although the genome of this patient was highly structurally re-arranged, the methylation

ratio of PC1 remained relatively stable and showed high fidelity in estimating tumour fraction (**Fig. 3.2.3.**) Although PC2 was mainly driven by only one patient with distinct genomic aberrations, this methylation signature raises the hypothesis that circulating methylation data could identify sensitivity to DNA damaging therapies.

## 3.8 Discussion - Challenges of accurate plasma methylome characterisation

### 3.8.1 Library construction and targeted enrichment

In my study, I constructed methylation libraries from bisulfite-converted low-input cfDNA in plasma. Then I applied a pre-designed target panel which aimed to capture over 5.5 million CpG sites. The panel was an expanded version of Infinium HumanMethylation450K BeadChip with extra coverage of adjacent CpGs. The post-capture libraries then underwent NGS on either HiSeq2500 or HiSeqX-10. There were some novel aspects to the wet lab workflow; 1) performing bisulfite conversion before library construction and 2) Targeted capture on libraries derived from low-input plasma DNA. However, it remains controversial if performing bisulfite conversion prior to library generation can significantly improve library complexity and quality. Since bisulfite conversion damages the DNA, break it into shorter fragments and make it single-stranded the first step of the library generation is to ligate single-stranded, truncated adaptor to the single-stranded DNA. This process is much less efficient than double stranded DNA ligation. Further experiments may be required to answer this question. For example, one can compare the DNA yield after bisulfite conversion and overall molarity of final library product. Also, WGBS saturation analyses on plasma samples could allow fair comparison of methylation library quality. However, all in all our selected approach appear to have generated biologically and clinically meaningful data.

### 3.8.2 Plasma methylome analysis workflow

The current analysis workflow of methylation library sequencing data consisted of read quality assessment, alignment against reference genome (hg19), removal of PCR duplicates, methylation ratio calling, as well as downstream analysis (discussed in the following section 6.3). Of all the steps, mapping of sequencing reads against reference genome using BSMAP [91] was the most time-consuming step and involved a significant computational burden. In the near future, if as predicted sequencing cost will keep dropping, WGBS data will be cheaper to obtain. It is thus crucial to solve this data processing bottleneck and improve the efficiency of WGBS data analysis.

Employing graphics processing unit (GPU), a micro-processer specialised in image rendering, can potentially solve this issue. In general, the GPU-based algorithm has demonstrated superior performance in deep learning especially for image processing to the CPU-based one. Although a CPU core can be more powerful, GPU is better in task parallelism. When it comes to huge tasks which includes multiple similar jobs, a GPU can help speed up the process. Multiple GPU-based genomic aligners such as BarraCUDA, CUDAlign or NextGenMap have been proposed to improve efficiency [116-119]. Recently, Arioc [120,121], a GPU-based aligner for bisulfite-converted sequencing reads, showed better mappability and alignment speed over CPU-based aligner such as Bismark [122]. The initial evidence showed that adapting a GPU-based algorithm can not only improve efficiency but also maximise the information we can collect from all sequencing reads.

### 3.8.3 Optimisation of methylation-based tumour fraction estimation

Here I characterised the plasma methylome in mCRPC and identified ct-MethSig whose methylation patterns were highly associated with tumour fraction. I used a custom target-capture approach to define the methylation status of pan-genome CpG islands. By using a 100bp sliding window strategy, I obtained close to 0.5 million methylation segments present in all of the 19 baseline plasma DNA samples and used these to construct the PCA. Novel to the methylation analysis was the construction of our model using solely mCRPC plasma DNA that comprised a variable ratio of normal DNA, primarily arising from white blood cells and tumour DNA that harboured methylation changes that were either prostate epithelium-specific or cancer-specific. By using the median methylation ratio of ct-MethSig, these findings could be generalised to different methods including methylation microarrays or reduced representation bisulfite sequencing with variable CpG coverage.

Ct-MethSig which spans 100+ CpG islands has the potential to sensitively track tumour changes at an early disease stage. However, since the signature was constructed on CRPC plasma by a pre-designed targeted panel, there might be some important methylation features missing. For example, the ct-MethSig prostate cancer specific segments could contain methylation events specific to castration resistance disease and potentially absent in the hormone-sensitive stage. Also, there are over 28 million CpG sites spanning across the genome, the targeted panel at best captures 5.5 million of them. It is very likely that there are other informative CpG sites missed in the current analysis. WGBS can solve this

issue, and as the sequencing technologies keep evolving, the cost of WGBS is expected to drop and becomes affordable for clinical implementation.

Moreover, the analysis performed herein used a window-based strategy by combining adjacent CpGs into a methylation segment and calculated the median methylation ratio of all CpG sites within the segment as a proxy of methylation level. This approach might introduce bias in segments where CpGs were not co-methylated and would fail to detect differentially variable CpG (DVC) which has been proven to be critical in normal physiological processes and other cancer types [123] [124]. Except for fix-length window, it is also feasible to use extract methylation ratios of all the CpG sites within a CpG island and take the median or average methylation ratio, or CpG density as a new feature.

I used principal component analysis to deconstruct the variability of plasma methylome which led to majority of the findings in my thesis. In general, methylation-based tissue signal deconvolution falls into two main categories – "reference-based" and "reference-free" approach. The former approach requires a good quality reference database. The plasma methylome from a mCRPC patient is mainly contributed by white blood cell methylome and prostate cancer methylome. If there is an existing, high-quality database that contains NGS deep sequencing from white blood cells and prostate cancer tissues, one can apply quadratic programming for tissue decomposition. Quadratic programming is a statistical process that performs linear regression and minimise several variables subject to normalisation constraints and can be used directly to estimate the

circulating tumour DNA fraction of a plasma sample. The latter approach aims to address the major confounder of the samples, and there are several mathematical methods such as surrogate variable analysis (SVA) or PCA, independent component analysis (ICA)[125] and non-negative factorization (NMF) available for *de novo* tissue decomposition. In my study, I employed PCA due to high quality data availability; however, there were some constraints that might limit the data interpretation and findings.

First, the main assumption was that circulating cell-free DNA tumour fraction being the major determinant of plasma methylome variance. The first component of PCA would be the one that best explains the variability of the data. The result may be misleading if the plasma samples subject to PCA did not have a wide range of tumour fraction because the major determinant for the variability may be from other tissues or white blood cell composition.

Second, each component of PCA is orthogonal to each other, but not as an independent component of the dataset. Thus, using PCA for tumour subtyping may also be challenging as two different components may still be highly correlated. In this case, ICA, which finds each vector as an independent component to the data, may be more suitable for this purpose.

Lastly, the PCA was currently built only on mCRPC plasma samples with median tumour fraction over 40% without other tissue samples such as lung or white blood cells. This may limit the usability of my findings in earlier disease stages, as

most CSPC plasma samples would have tumour fraction less than 15% and normal tissue contaminations may be too significant to ignore.

# 4 Chapter 4. Implementation of a methylation signature for tracking and detection of prostate cancer in plasma

## Hypotheses

1. It is possible to build a classifier to identify plasma DNA derived from cancer patients.
2. Feature selection may help improve detection sensitivity

## Aims

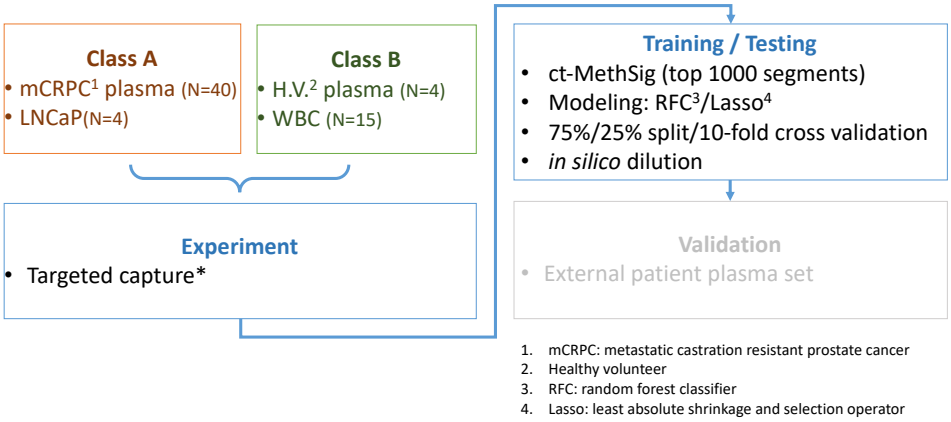
1. To build a classifier using classic machine learning methods such as random forest classifier (RFC) or LASSO.
2. To adapt pre-selected features (ct-MethSig prostate-specific and/or cancer-specific segments) as inputs for machine learning model.

## 4.1 Prostate cancer detection using plasma methylome

The clinical unmet need is to identify clinical aggressive forms of the disease which actively sheds DNA into the circulation. Here I aimed to build a classification model to predict plasma samples containing circulating cell-free DNA derived from prostate tumour cells. I have successfully identified that in prostate cancer patients the main contributor of plasma DNA methylation variance was from prostate cancerous tissues. The methylation signature (ct-MethSig), of which the methylation levels can be used as a proxy for tumour fraction, comprised of prostate tissue specific and prostate cancer specific methylation patterns. This information was crucial for building a classification model.

I used the metastatic prostate cancer plasma samples (N = 44) as described before (**Table 2.1.1.**) plus fifteen leukocyte samples derived from patients and two healthy volunteer plasma and leukocyte samples. I labelled the patient plasma samples as class A while the leukocyte and samples collected from healthy volunteer as class B. The goal was to build a classifier to accurately categorise class A and class B (**Figure 4.1.1.**).

Fig. 4.1.1.
Workflow of building a classification model




1. mCRPC: metastatic castration resistant prostate cancer
2. Healthy volunteer
3. RFC: random forest classifier
4. Lasso: least absolute shrinkage and selection operator

The methylation ratios of ct-MethSig across all samples were used as input for random forest classifier (RFC), a machine learning classification method. A RFC model was built on and fitted a number of decision trees each of which categorized a subset of samples to improve the prediction accuracy and control for overfitting. The RFC was run with 1000 times cross-validation to ensure the stability of the model. In short, the samples were split into two groups – a training group and a testing group. The classification model was initially built on the training group and the classifier was tested on the testing group. I initially started to build the model selecting 10 trees in one forest, and the result showed 100% accuracy (STD = 1%) on training and 95% on testing (STD = 11%, **Figure 4.1.2**.). When I increased the number of trees in the forest to 100, the model performance slightly improved to 100% accuracy (STD = 1%) on training and 97% on testing (STD = 9%, **Figure 4.1.3**.).

Fig. 4.1.2.
Accuracy of Random Forest Classification model (number of trees in the forest = 10) on 1000-time cross validation

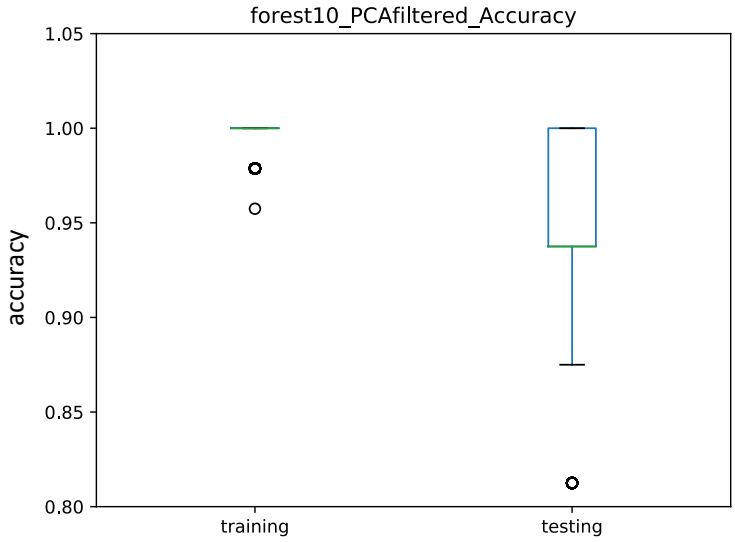
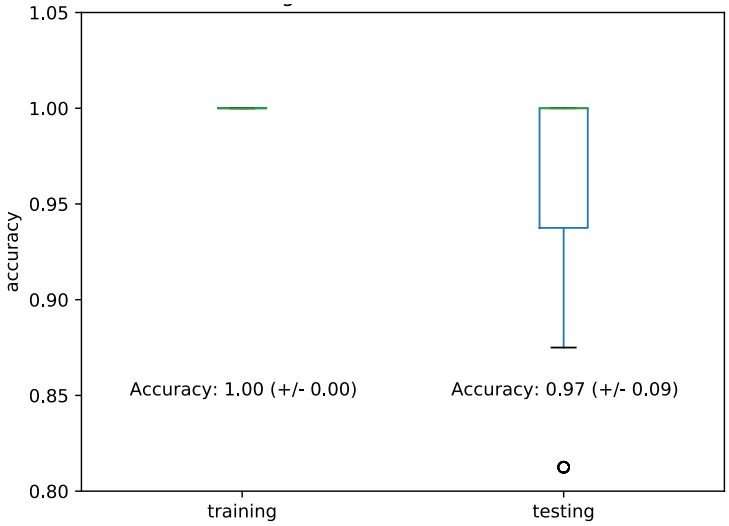



Fig. 4.1.3.
Accuracy of Random Forest Classification model (or RFC, number of trees in the forest = 100) on 1000-time cross validation



I also evaluated the least absolute shrinkage and selection operator (LASSO) method, a regression analysis algorithm popular for variable selection, on the same data input as described above. The LASSO method was also subject to 100-time cross validation to confirm the model stability. As a result, the training and testing accuracy of LASSO with $\alpha$ value 0.01 were 77% (STD = 9%) and 66% (STD = 14%) respectively (**Figure 4.1.4**.). The LASSO model ($\alpha$ = 0.01) reduced the feature number down to nine segments (**Table 4.1.1.**). When I used a smaller alpha value such as 0.0001, the training accuracy improved to 100% (STD = 1%) and the testing accuracy was 71% (STD = 18%) and the number of methylation features used were thirty-three (**Table 4.1.2**.). The LASSO method tended to overfit the training dataset and the accuracy generally performed worse than the RFC method (**Figure 4.1.5**.).

Table. 4.1.1. List of segments used for LASSO model ($\alpha$ = 0.01)

| chr | start | end |
|---|---|---|
| chr10 | 120006301 | 120006401 |
| chr10 | 7449701 | 7449801 |
| chr2 | 11496951 | 11497051 |
| chr2 | 3246251 | 3246351 |
| chr5 | 72683901 | 72684001 |
| chr6 | 19692251 | 19692351 |
| chr6 | 26172301 | 26172401 |
| chr6 | 26189301 | 26189401 |
| chr6 | 34203851 | 34203951 |

Table. 4.1.2. List of segments used for LASSO model ($\alpha$ = 0.0001)

| chr | start | end |
|---|---|---|
| chr1 | 119548401 | 119548501 |
| chr1 | 39991501 | 39991601 |
| chr10 | 120006301 | 120006401 |
| chr10 | 4125351 | 4125451 |
| chr11 | 122722201 | 122722301 |
| chr11 | 46298351 | 46298451 |
| chr12 | 104526451 | 104526551 |
| chr12 | 130936451 | 130936551 |
| chr12 | 54409101 | 54409201 |
| chr12 | 6756551 | 6756651 |
| chr12 | 75728001 | 75728101 |
| chr14 | 102551401 | 102551501 |
| chr14 | 104668851 | 104668951 |
| chr15 | 37330151 | 37330251 |
| chr15 | 38670451 | 38670551 |
| chr15 | 88360201 | 88360301 |
| chr15 | 88360251 | 88360351 |
| chr15 | 96913151 | 96913251 |
| chr17 | 81047501 | 81047601 |
| chr19 | 43979601 | 43979701 |
| chr2 | 128453451 | 128453551 |
| chr2 | 177022251 | 177022351 |
| chr20 | 9489851 | 9489951 |
| chr21 | 39870351 | 39870451 |
| chr21 | 43183101 | 43183201 |
| chr3 | 186193951 | 186194051 |
| chr4 | 157682751 | 157682851 |
| chr5 | 5033251 | 5033351 |
| chr5 | 5033301 | 5033401 |
| chr6 | 19692251 | 19692351 |
| chr8 | 42037251 | 42037351 |
| chr8 | 42037451 | 42037551 |
| chr9 | 35729701 | 35729801 |

Fig. 4.1.4.
Accuracy of LASSO model ($\alpha$ = 0.01) on 100-time cross validation

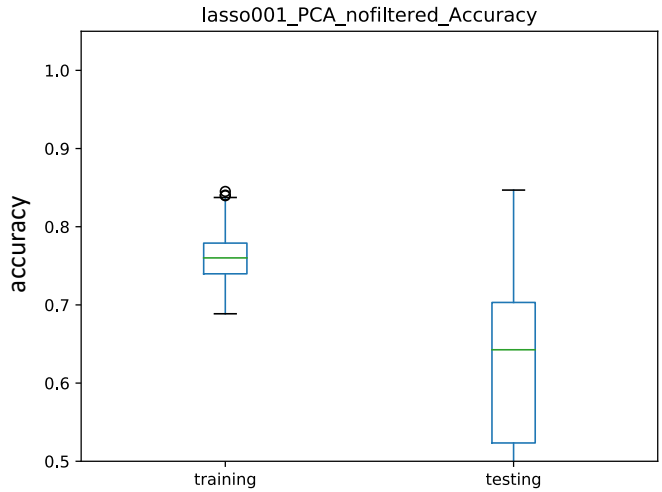
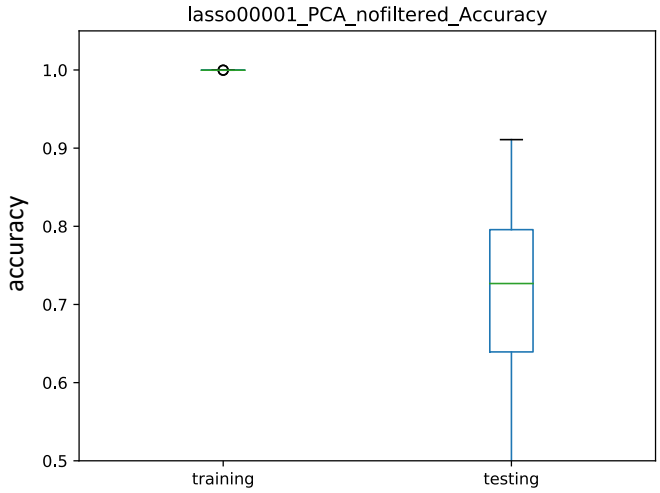


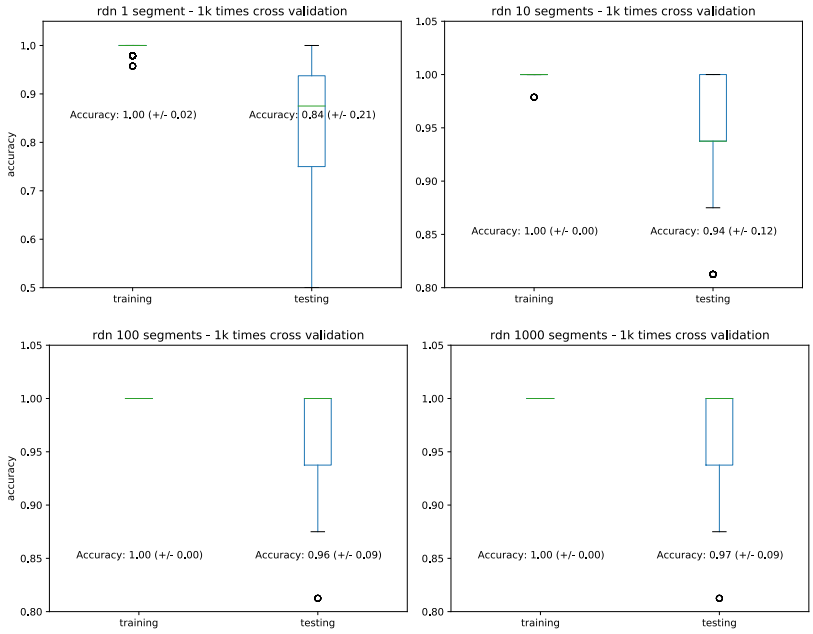

Fig. 4.1.5.
Accuracy of LASSO model ($\alpha$ = 0.0001) on 100-time cross validation




Also, I was interested to investigate whether the randomly selected 1, 10 or 100 segments would be enough to construct a reliable classifier. Therefore, I randomly

selected a fixed number of segments (1, 10, and 100), and used these segment(s) to build RFC (n_estimators = 100) with 1000-time iteration. The results indicated that using only one randomly selected the testing accuracy was 84% (STD% = 20%). The testing accuracy gradually improved when I included more segments (**Figure 4.1.6**).

Fig. 4.1.6.
Accuracy of RFC model (number of trees in the forest = 100) on 1000-time cross validation trained on 1, 10, or 100 randomly selected ct-MethSig segments

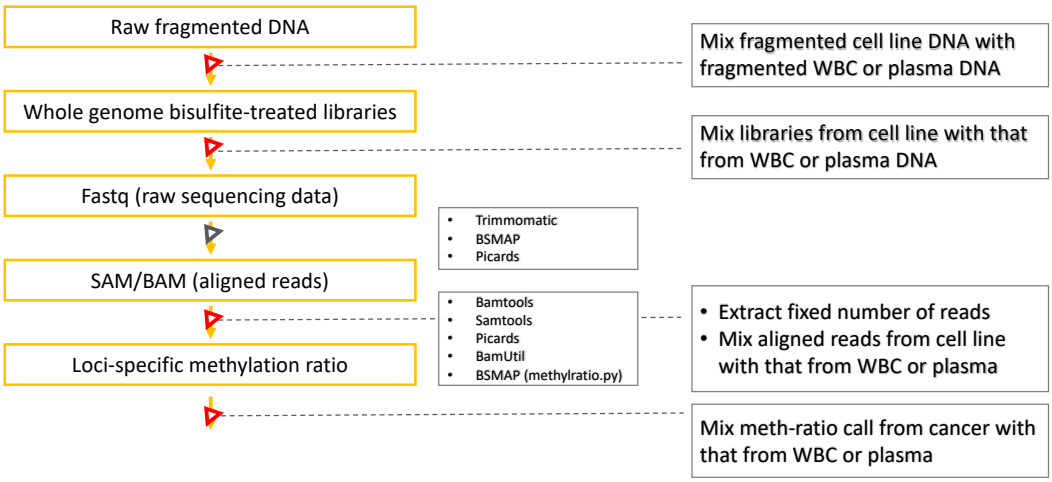



In summary, the development of a methylation based classifier was achievable and able to identify plasma samples containing circulating tumour DNA with high accuracy. The RFC approach seemed to outperform LASSO method in the dataset with higher testing and training accuracy and was less likely to overfit the training dataset.

## 4.2 Detection positivity of the classification model

To test the detection limit of the RFC model, I aimed to perform in silico dilution to define the test detection sensitivity. In principal, there are many ways of performing the dilution experiment in order to properly define the assay limitation (**Figure 4.2.1.**). First, one can mix sonicated cell line or pure tumour DNA with white blood cell DNA or plasma DNA obtained from healthy volunteers. This approach is less preferable because during the library generation the plasma DNA subjected to bisulfite treatment usually has much more amplifiable library molecules than artificially fragmented DNA. This phenomenon results in the final library more enriched in sequenceable DNA molecules from healthy volunteer than from tumour cell line or tissue, and thus the actual tumour fraction would be much lower than the estimation. Secondly, dilution experiment can be done by mixing libraries derived from cell line or cancer tissue with libraries from healthy volunteer plasma or white blood cell. However, to create a gradient of tumour fraction, multiple indexes are required, and the libraries from the same tissue source may need to be prepared separately. This is less feasible for plasma DNA as the amount is usually limited.

Fig. 4.2.1.
Workflow of methylation analysis and dilution experiment to define assay sensitivity




In silico serial dilution seems to be a more reliable and less variable approach. It is suggested that the dilution experiment can be done by down-sampling reads before or after read-clipping and mix reads from cancerous tissue with reads from healthy volunteer or non-cancerous tissue. One potential downside of doing so is the coverage may drop during the down-sampling procedure and makes the methylation ratio calling less accurate. Last but not least, as methylation status is a stable and quantifiable epigenetic marker, serial dilution can be executed by using weighted sum of methylation level from cancerous tissue and non-cancerous tissue.

I executed in silico dilution by using weighted sum of methylation ratio derived from LNCaP and the healthy volunteer plasma sample. The diluted samples were then subjected to the RFC model as described previously with 100-times cross

validation to test the detection limit. The detection limit was reported as the positivity percentage. For example, if a model was run 100 times and 20 out of 100 times this model reported a positive result, the positivity would be 20%. On an in-silico dilution sample of 4% tumour fraction, the RFC model (number of trees in the forest = 100) predicted the sample 88 times out of 100 to be positive of tumour (**Figure 4.2.2**.). On a 3% tumour fraction sample (in-silico dilution), the model only predicted the sample 45 times out of 100 to be positive. I further increased the number of trees in the RFC model and the modified model showed better detection positivity (**Figure 4.2.3**.). Furthermore, since the methylation level of ct-MethSig negatively correlated segments was more accurately correlated with genomically-determined tumour fraction, I hypothesized that using ct-MethSig negatively correlated segments could improve detection sensitivity. I used methylation level of ct-MethSig negatively correlated segments only as inputs for RFC model and the detection positivity showed that, on the 2% tumour fraction sample, the model was able to detect tumour 64 out of 100 times (**Figure 4.2.4**.**A**). However, the improved detection positivity was compromised by the decrease in testing accuracy of the model (**Figure 4.2.4**.**B**). This could be due to overfitting as the feature number decreased. Moreover, selecting prostate cancer-specific segments or normal prostate epithelium-specific segments did not improve the RFC model detection sensitivity (**Figure 4.2.5**.). However, the use of normal prostate epithelium-specific segments in a model might help distinguish benign prostate hyperplasia or prostatitis from prostate malignancy.

Fig. 4.2.2.
RFC model (number of trees in the forest = 100) detection sensitivity

| tumour fraction | positivity ratio % |
|---|---|
| 0.50% | 0 |
| 1% | 0 |
| 2% | 11 |
| 3% | 45 |
| 4% | 88 |
| 5% | 100 |
| 10% | 100 |
| 20% | 100 |
| 50% | 100 |

Fig. 4.2.3.
RFC model (number of trees in the forest = 1000) detection sensitivity

| tumour fraction | positivity ratio % |
|---|---|
| 0.50% | 0 |
| 1% | 1 |
| 2% | 12 |
| 3% | 52 |
| 4% | 93 |
| 5% | 100 |
| 10% | 100 |
| 20% | 100 |
| 50% | 100 |

Fig. 4.2.4.
(A) RFC model (number of trees in the forest = 100) detection sensitivity (using ct-MethSig positively correlated segments only)
(B) Accuracy of the RFC model

(A)

| tumour fraction | positivity ratio % |
|---|---|
| 0.50% | 16 |
| 1% | 26 |
| 2% | 64 |
| 3% | 87 |
| 4% | 97 |
| 5% | 100 |
| 10% | 100 |
| 20% | 100 |
| 50% | 100 |

(B)

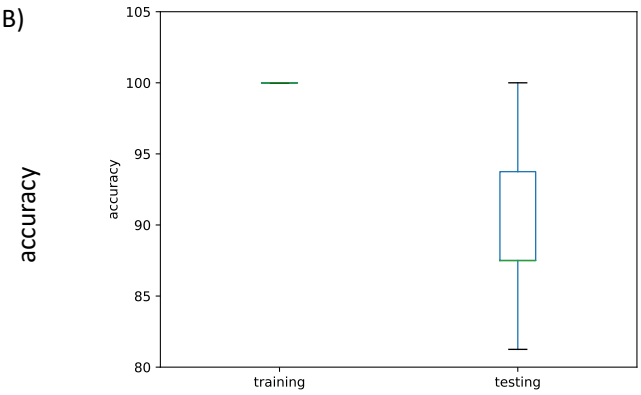


Fig. 4.2.5.
(A) RFC model (number of trees in the forest = 100) detection sensitivity (using ct-MethSig prostate cancer-specific segments only)
(B) RFC model (number of trees in the forest = 100) detection sensitivity (using ct-MethSig prostate epithelium-specific segments only)

(A)

| tumour fraction | positivity ratio % |
|---|---|
| 0.50% | 15 |
| 1% | 18 |
| 2% | 31 |
| 3% | 66 |
| 4% | 87 |
| 5% | 98 |
| 10% | 100 |
| 20% | 100 |
| 50% | 100 |

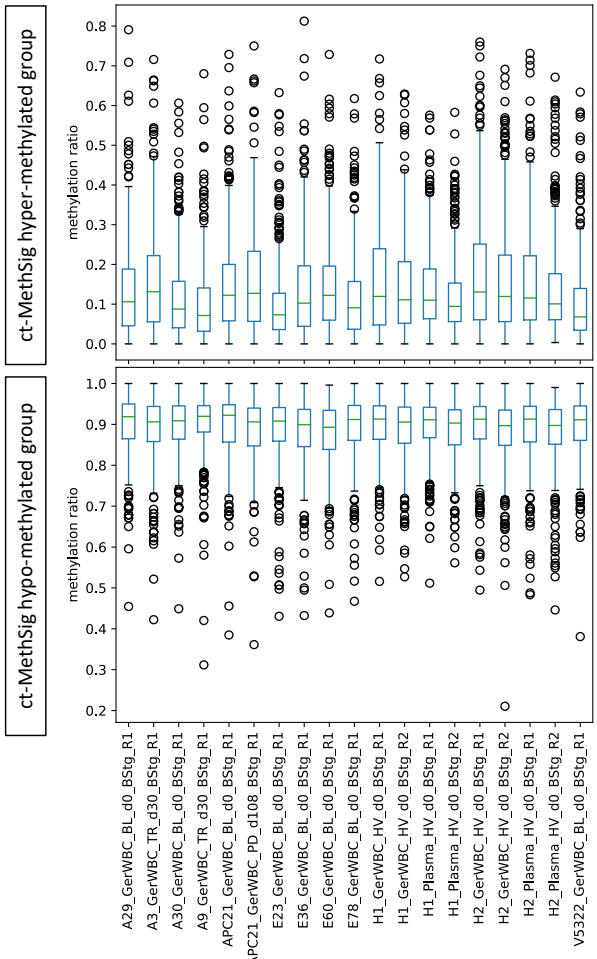
(B)

| tumour fraction | positivity ratio % |
|---|---|
| 0.50% | 12 |
| 1% | 15 |
| 2% | 33 |
| 3% | 67 |
| 4% | 91 |
| 5% | 98 |
| 10% | 100 |
| 20% | 100 |
| 50% | 100 |

To assess detection limitation, I plotted the methylation ratio of ct-MethSig of white blood cell and healthy volunteer plasma (**Figure 4.2.6**.). The observed inter-sample and inter-individual variability may explain the current detection limitation of RFC model. To conclude, a clinically-applicable classification model was able to accurately identify plasma tumour sample. A tree-based algorithm on

methylation data analysis along with the feature selection based on principal component analysis tended to outperform a classic linear model and avoid overfitting. The inter-individual and technical variabilities may be the hurdle to further improve the detection sensitivity (see **Chapter 6 Discussion**).

Fig. 4.2.6.
ct-MethSig methylation ratio of white blood cells and healthy volunteer plasma



## 4.3 ct-MethSig in castration-sensitive prostate cancer plasma samples

The presence of circulating tumour DNA following local, curative treatment has shown to be linked with worse clinical outcome in many cancer types including lung, breast and colorectal cancer. In prostate cancer, it remains unclear whether detection of circulating tumour DNA would be associated with more aggressive clinical courses. Under the current practice, there are some clinical windows to apply ct-MethSig for early detection of disease relapse. For example, ct-MethSig could guide clinicians to intensify treatment in patients with high risk, localised prostate cancer who would receive local curative therapy followed by ADT, or in patients on long-term ADT with or without clear metastatic signs (M1 or M0 HSPC). I hypothesized that MRD at the hormone-sensitive stage can be detected using the prediction model built on ct-MethSig (**Chapter 4.2.**). Also, ct-MethSig could potentially be used alone or in combination with PSA level to identify patients which may benefit from additional treatments.

As a start, I applied ct-MethSig which was built on mCRPC patients as described before (Chapter 3) on plasma samples collected after the start of ADT at HSPC. The samples were subjected to targeted methylation analysis and methylation ratio of all on-target segments, including ct-MethSig segment, were extracted. As described before, the methylation ratio of the Top 1000 PC1 correlated segments can be used as a proxy for tumour fraction, I thus plotted the ct-MethSig negatively and positively correlated segments of all HSPC plasma samples (**Fig.**

**4.3.1.**). It was obvious that some plasma samples may contain circulating tumour DNA as the methylation ratio across ct-MethSig deviated from that of healthy volunteers. Later I applied the eigenvectors used to construct principal component analysis based on all mCRPC plasma samples to calculate principal component one value of each HSPC plasma sample which can be used to estimate tumour fraction. Also, I employed the prediction model described in the previous section (**Chapter 4.2.**) to predict the likelihood of each HSPC sample containing tumour (**Table. 4.3.1**).

Fig. 4.3.1.
ct-MethSig methylation ratio of CSPC plasma samples, normal prostate, and healthy volunteer plasma

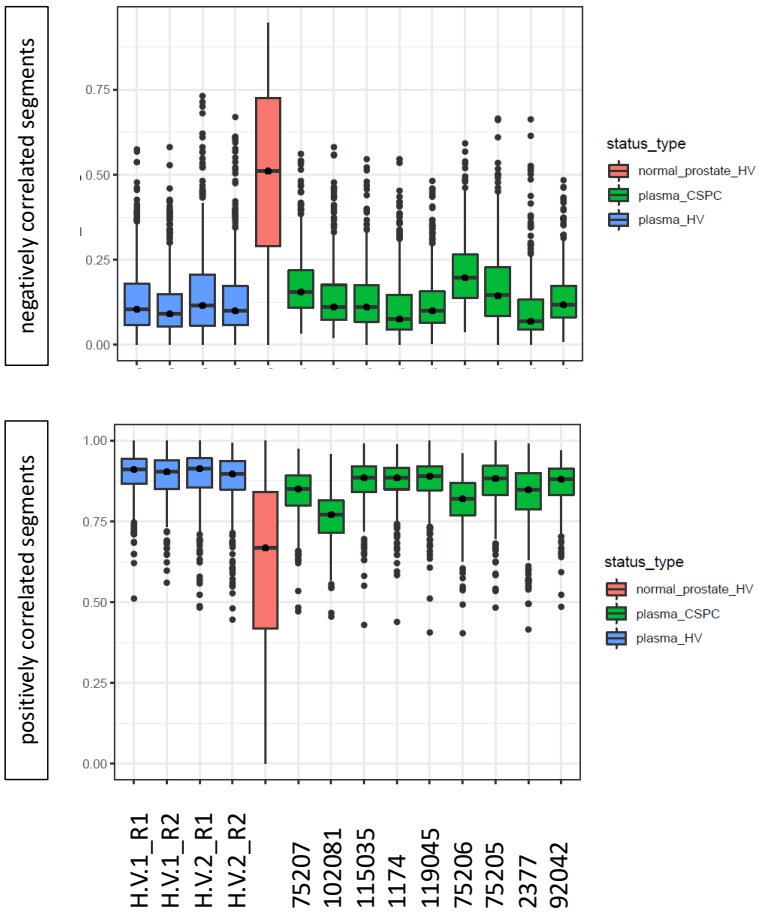


Table. 4.3.1.
CSPC plasma samples (from STAMPEDE trial)

| Trial ID | Mts status | Collection Date | ADT start date | Time from start ADT and collection date | Type of ADT | tumour fraction | ct-MethSig detection positivity |
|---|---|---|---|---|---|---|---|
| 75207 | M1 | 12/02/2018 | 28/12/2017 | 46 | LHRH antagonist | 7% | 98.40% |
| 102081 | M1 | 21/02/2018 | 04/01/2018 | 48 | LHRH agonist | 17% | 100% |
| 115035 | M1 | 08/03/2018 | 01/03/2018 | 7 | LHRH agonist | 1% | 2% |
| 1174 | M1 | 13/03/2018 | 10/01/2018 | 62 | LHRH agonist | 0% | 0% |
| 119045 | M0 | 02/02/2018 | 21/12/2017 | 43 | LHRH agonist | 0% | 1% |
| 75206 | M1 | 13/02/2018 | 13/12/2017 | 62 | LHRH antagonist | 13% | 100% |
| 75205 | M1 | 16/02/2018 | 23/01/2018 | 24 | LHRH agonist | 0% | 9% |
| 2377 | M0 | 19/02/2018 | 08/02/2018 | 11 | LHRH agonist | 5% | 80.50% |
| 92042 | M0 | 20/02/2018 | 02/01/2018 | 49 | LHRH agonist | 3% | 60% |

Even though all HSPC plasma samples were collected after the start of androgen deprivation therapy (i.e. LHRH agonist or LHRH antagonist), the result still indicated that at least three out of nine samples contained tumour. These three patients (trial ID: 75206, 75207, and 102081) all harboured metastatic disease, and two out of three patients (trial ID: 75206 and 75207) were treated with LHRH

antagonist, which was probably chosen for patients with suspected higher disease burden. This preliminary result, albeit limited by the number of patients and lack of long-term clinical follow-up, showed that the plasma methylome can be used to detect tumour at earlier stages. Although exposure to ADT at HSPC stage is expected to control disease effectively at least initially, there were a few cases which had detectable circulating tumour signatures after treatment initiation (See **4.4 Discussion**).

## 4.4 Discussion - Applications and challenges of methylation-based ctDNA detection

In keeping with recent studies, screening for circulating tumour DNA based on methylation data [13,14,16,17,19,79,80] is a potentially sensitive and accurate approach for tracking tumour dynamics. In metastatic disease the ctDNA fraction has been linked with tumour loads, while in earlier disease stages, ctDNA fraction can be a sensitive indicator to detect minimal residual disease (MRD) [70,71,84]. Studies in other tumour types have used genomic ctDNA analysis to detect MRD. For example, Abbosh et al. examined pre- and post-surgery plasma DNA in a group of patients (N=23) and found that 13 patients tested positive for post-operative plasma ctDNA, all of whom relapsed. Conversely, 9 out of 10 patients who tested negative for ctDNA remained disease-free [71]. In colon cancer, preliminary data has also indicated that detection of MRD in patients with resected stage II colon cancer had worse outcome [19]. In this study of 230 patients with stage II diseases, in patients not treated with adjuvant chemotherapy, ctDNA was detected postoperatively in 14 of 178 (7.9%) patients, 11 (79%) of whom had recurred at a median follow-up of 27 months while recurrence occurred in only 16 (9.8 %) of 164 patients with negative ctDNA.

Based on the methylation data derived from CRPC patients, I built a classification model to stratify plasma samples into a ctDNA positive group and a negative group. The initial results revealed that the classifier can identify ctDNA positive samples collected from on treatment (i.e. anti-androgen therapy, or ADT) CSPC patients.

This result was encouraging as I showed that the model can be applied to the hormone-sensitive stage. However, challenges remain such as inter-individual variations in methylation patterns of normal tissues such as white blood cells which could be hard to ignore given the lower tumour fraction at the CSPC stage. Also, other machine learning models such as XgBoost should be tested, either alone or in combination with the random forest classification model, to see if the classification accuracy can be improved.

Furthermore, instead of using a single CpG or methylation segment methylation ratio as an input, CpG island (CGI) or methylation haplotype load can also be considered. Recently, Guo et al. proposed that methylation haplotype loads performed better in tissue classification based on methylation data than single CpG methylation ratio or weighted methylation ratio across multiple adjacent CpGs [13]. Last but not least, an external validation set with complete clinical follow-up data would be necessary to justify the clinical implementation of the methylation-based classification model. Some ongoing and future clinical trials have been designed to serve this purpose – to detect and track disease changes at earlier stages.

# 5 Chapter 5. Methylation signatures specific to CRPC

## Hypotheses

1. Plasma methylome may contain biologically relevant and clinically useful information.
2. It is feasible to subtype tumours based on plasma methylation status.

## Aims

1. To extract subtyping methylation signatures, independent of tumour fraction.
2. To understand the biological consequences of methylation signatures.

## 5.1 Methylation signatures specific to an individual's cancer

I later focused on the third principal component (PC3) which contributed to 8% of global plasma methylome variance. The principal component showed a weak correlation with tumour fraction (r=0.01, *P* = 0.96, Pearson correlation). Similar to the methodology applied to ct-MethSig, I first identified the top 1000 segments that were most correlated with this component's values. In contrast to ct-MethSig, these were predominantly positively correlated (**Fig. 5.1.1.**). Using the median of every segment's methylation ratio, I was able to incorporate array-based methylation data from biopsies from intermediate-risk HSPC[38] and mCRPC[52]. I found that the median methylation ratio in CRPC plasma and tumour samples presented a higher variance in contrast to intermediate-grade HSPC or white blood cells (**Fig. 5.1.2. and 5.1.3.**). Also, I plotted the PC3 values against tumour fraction and found that the values remained relatively stable and showed little intra-patient variability (**Fig. 5.1.4.**). In contrast to ct-MethSig, I confirmed that a change in tumour fraction before and after treatment did not change the median methylation ratio of the top correlated segments with principal component 3 (**Fig. 5.1.5.**). Similarly, inter-patient differences were greater than intra-patient variability in multiple metastases harvested from the same patient at autopsy (**Fig. 5.1.6.**). To further leverage the CASCADE rapid warm autopsy samples and evaluate whether plasma DNA shows the same methylation signatures as metastatic sites, I performed whole genome bisulfite sequencing on plasma DNA from the 4 (CA27, CA34, CA35, and CA43) men it was available for. The data shown in **Fig. 5.1.6.** indicated that the AR-MethSig score in plasma samples matches that of metastatic tissues with less intra- than inter-patient variability.

Fig. 5.1.1.
Top 1000 segments with highest correlation coefficient between the third principal component (PC3) and methylation ratio

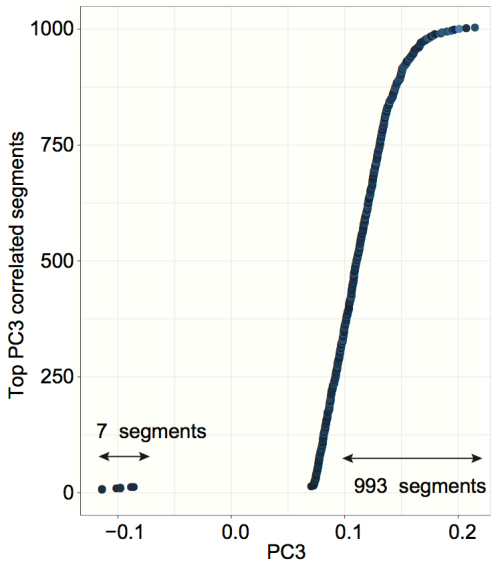



Fig. 5.1.2.
Median methylation ratio of 993 segments positively correlated with PC3 values across different sample types—plasma, white blood cells, cell lines (LNCaP, LNCaP95, VCaP), CASCADE tumour (mCRPC biopsy) are plotted against the median methylation ratio of top correlated segments with ct-MethSig.

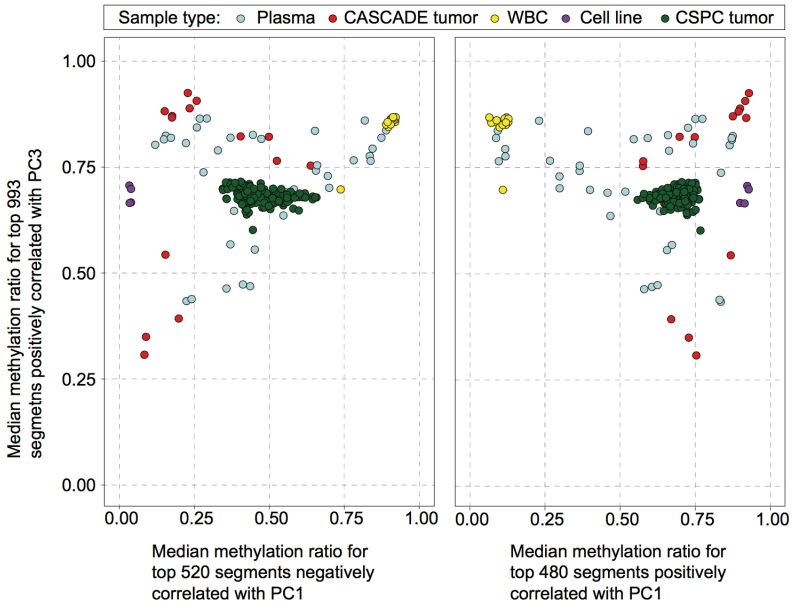


Fig. 5.1.3.
Methylation ratio of top 1000 segments highly correlated with PC3 values derived from plasma, white blood cell, HSPC tumour, and CRPC tumour (CASCADE trial)

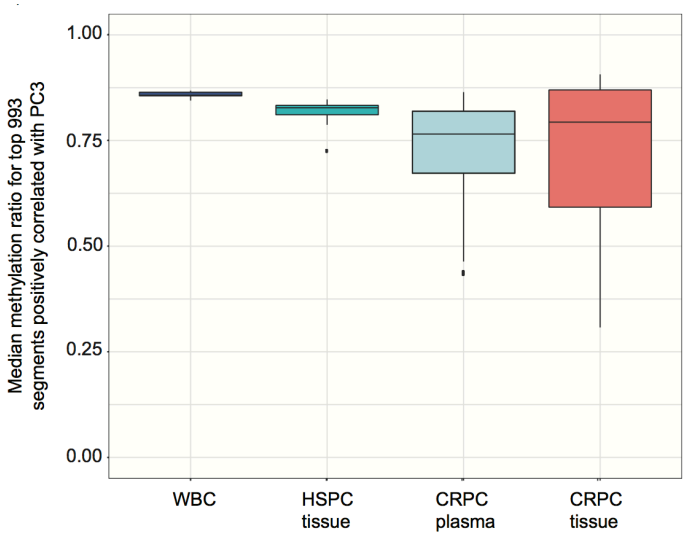


Fig. 5.1.4.
Principal component three values plotted against tumour fraction by sample

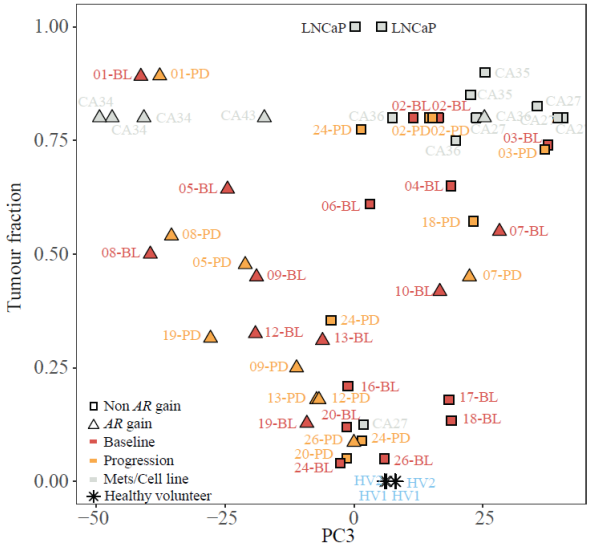



Fig. 5.1.5.
Comparison of intra-individual changes in the top correlated segments defined by targeted methylation NGS on plasma DNA and changes in tumour fraction. Y-axis denotes the difference (Δ) of mean methylation ratio of the top correlated segments between baseline and progression samples and the X-axis denotes the difference (Δ) in tumour fraction.

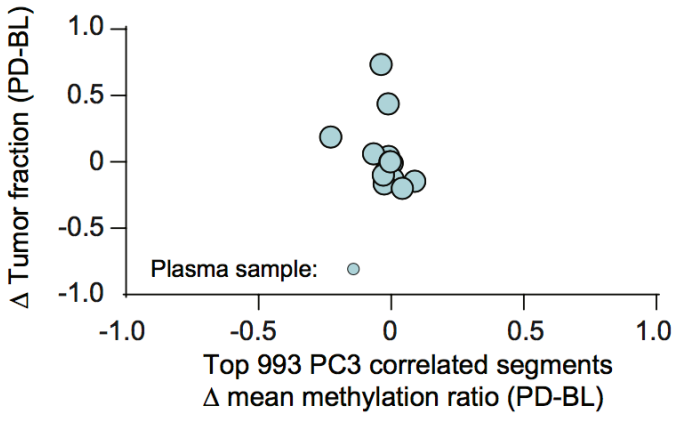
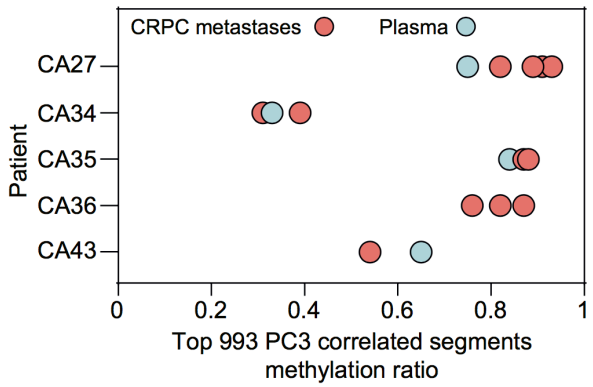


Fig. 5.1.6.
Median methylation ratio of the top correlated segments of different metastatic sites by patient from the CASCADE rapid warm autopsy program.



## 5.2 Enrichment analysis on the PC3 top 1000 correlated segments identified AR-binding motif

As I did for ct-MethSig, I performed gene set enrichment and pathway analysis to identify commonly regulated pathways, and the analysis on the PC3 top 1000 correlated segments showed enrichment in histone H3 tri-methylation marker (H3K27Me3, **Table. 5.2.1. and Table 5.2.2.**) similar to the finding for PC1. The finding indicated that the methylation event was primarily juxtaposed with the histone modification and also confirmed the dynamic interplay between histone repressive epigenetic markers and DNA methylation. I hypothesized that this methylation signature (PC3 top correlated segments) could be regulated by a common transcriptional pathway and that transcription factor binding to these segments introduced variance in methylation levels. I therefore searched for known transcriptional factor binding sites (TFBSs) within 75 base-pairs of the start of the top 1000 segments using a protocol described previously [110]. Notably, the AR binding motif was the only significantly over-represented binding site (local enrichment $P = 6X10^{-4}$, global enrichment $P = 3X10^{-16}$; **Fig. 5.2.1., Supplementary Table. 5.2.3.**). Hence, I denoted this profile as AR-MethSig.

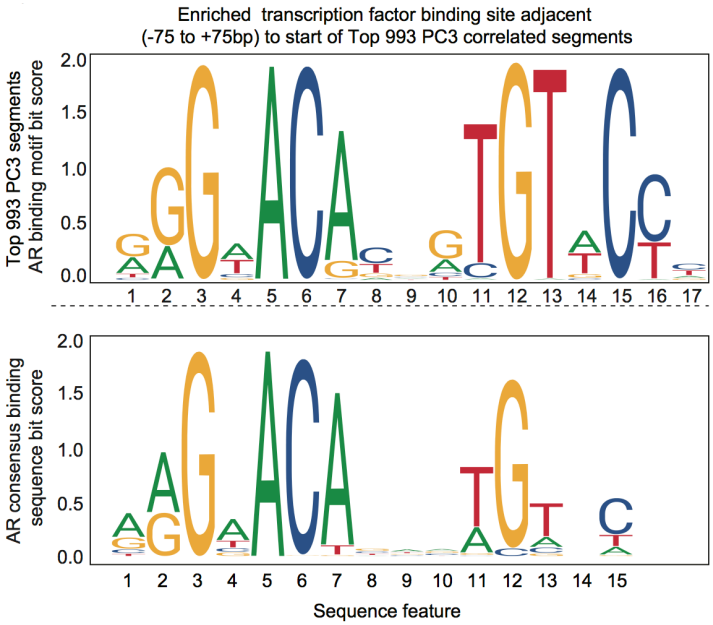
Table. 5.2.1.
Functional enrichment of principal component three top 1000 segments

| Term | Type | ID | Input.numbe | Background.r | P.value | adjusted.P.value |
|---|---|---|---|---|---|---|
| MIKKELSEN_MCV6_HCP_WITH_H3K27ME3 | MSigDB | M1954 | 28 | 431 | 3.58E-11 | 1.19E-07 |
| MARTENS_TRETINOIN_RESPONSE_UP | MSigDB | M2098 | 35 | 706 | 1.64E-10 | 2.71E-07 |
| BENPORATH_ES_WITH_H3K27ME3 | MSigDB | M10371 | 37 | 995 | 1.08E-07 | 0.000119034 |
| BENPORATH_EED_TARGETS | MSigDB | M7617 | 34 | 921 | 5.45E-07 | 0.000451695 |
| BENPORATH_SUZ12_TARGETS | MSigDB | M9898 | 32 | 924 | 5.37E-06 | 0.003561478 |
| MEISSNER_NPC_HCP_WITH_H3K4ME3_AND_H3K27ME3 | MSigDB | M1935 | 10 | 144 | 2.96E-05 | 0.016353513 |
| NAKAMURA_METASTASIS_MODEL_DN | MSigDB | M15940 | 5 | 39 | 4.58E-05 | 0.01899284 |
| MEISSNER_NPC_HCP_WITH_H3_UNMETHYLATED | MSigDB | M1936 | 20 | 487 | 4.16E-05 | 0.01899284 |
| MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3 | MSigDB | M1941 | 32 | 1035 | 5.34E-05 | 0.019656077 |
| BENPORATH_PRC2_TARGETS | MSigDB | M8448 | 22 | 594 | 8.33E-05 | 0.027602582 |

Table. 5.2.2.
Genes overlapping with AR-MethSig

| | | | |
|---|---|---|---|
| WNT16 | NPY | CBS | TUSC5 |
| MEOX1 | CNTFR | SPTBN4 | SORCS2 |
| ASB4 | SARDH | PGLYRP2 | ADARB2 |
| FYN | PREX1 | DMKN | ADAMTSL5 |
| ZBTB32 | PTGIS | JAK1 | KLK12 |
| MAN2B2 | KCNK15 | MEGF6 | MPPED1 |
| INSRR | RUNX2 | DISC1 | TMPRSS6 |
| ELN | HS3ST3B1 | PQLC3 | SAMD11 |
| HEXB | TMEM74B | PF4 | ERC2 |
| FSTL4 | RFPL3 | ABLIM2 | GABRD |
| FOXN3 | CBFA2T3 | ANKRD33B | DNAH17 |
| HHAT | ZSCAN10 | HIST1H2AA | LRRIQ4 |
| CAMK2B | EXOC3L2 | AQP3 | PDCD1 |
| DGKG | COL5A1 | ELFN2 | KRTDAP |
| TLE2 | KHDRBS3 | MS4A8 | LINC00523 |
| SLC9A3 | NREP | PRRX2 | AJAP1 |
| KIF26A | SRPK2 | C16orf92 | ANXA6 |
| RORA | SCRN1 | ZNF180 | LINC00336 |
| SPTB | IGF2BP3 | LY6D | DNM3 |
| CNGB1 | LMX1B | C11orf85 | DLGAP2 |
| ST6GALNAC2 | MGARP | DEGS2 | CARD11 |
| EPHA8 | MAP2K5 | KLHL30 | NTRK1 |
| IGF2BP2 | ITGA11 | SFTPC | ZNF583 |
| PAG1 | BCAR3 | KCTD19 | C2CD4A |
| PKD2L2 | CDK15 | PCSK9 | SLC34A3 |
| CRYBG3 | SLC38A4 | WNT10B | SMOC1 |
| OPRK1 | GALNS | LRRN2 | RASSF9 |
| PILRA | KSR1 | MGMT | MUC2 |
| TGFB2 | ARSG | KSR2 | C1orf95 |
| BAMBI | ASGR1 | DSCAM | IGFL2 |
| HPS4 | MMEL1 | P2RY6 | PCDHGA1 |
| CACNA1I | PRDM16 | PLEKHG5 | EXOC3L4 |
| PVALB | FHAD1 | CAMTA1 | MIR548D2 |
| RIN3 | SUSD4 | NXNL1 | MIR133A2 |
| ASB2 | KCNN3 | FBXL14 | ARL2 |
| NTSR1 | MEIS1 | PRND | MIR1268A |
| CHRNA4 | GULP1 | LRRC15 | EBF2 |
| CCM2L | PTH2R | RCAN2 | PLXNA4 |
| MGRN1 | IQSEC1 | DAB1 | URAHP |
| PLLP | SLC25A26 | C2orf70 | LINC00703 |
| ZNF423 | NKD2 | SLC6A19 | LINC00162 |
| WFDC1 | ANKRD31 | LEP | LINC00705 |
| FAM189A1 | SLC17A4 | AMZ1 | ELFN1 |
| CGB | HIST1H2BA | GPR152 | MROH5 |
| ZFR2 | ANO7 | CABP4 | STEAP2-AS1 |
| BBC3 | SLC2A12 | LINC00521 | LINC00704 |
| ZNRF4 | C7orf50 | DLEU1 | LINC00689 |
| COMP | SDK1 | KBTBD11 | EMBP1 |
| VIPR2 | NTMT1 | OR51F2 | ADAM6 |
| CPVL | PARD3 | UMODL1 | DPY19L2P4 |
| CHN2 | SERPING1 | C2orf73 | ERICH1-AS1 |
| CRHR2 | GRIK4 | UCN3 | TDGF1 |
| CLIP2 | GGTLC1 | C9orf50 | MICAL3 |
| CDH23 | PLCH2 | HTR1D | ETV5 |
| EBF3 | SCN2B | TH | LINC00535 |
| RGS9 | CDH22 | PAK2 | FMN1 |
| SOD3 | NUDT22 | C9orf139 | GPR162 |
| FAM149A | CCDC3 | FUT7 | KBTBD11-OT1 |
| DNAJC4 | VENTX | AATK | CCDC177 |
| BIRC2 | TRIM36 | CCDC172 | MIR548W |
| CALCA | PITPNC1 | CAMK1D | ESPNP |
| SLC6A12 | FGD5 | URAD | TRABD2B |
| FGF1 | ODF1 | ASCL2 | TSNAX-DISC1 |
| RBP1 | KCNMA1 | B3GALT5 | |
| EFCC1 | CACNA2D3 | SLC35F3 | |
| PLCD4 | MEGF11 | ARSI | |
| NR5A2 | RADIL | TBX1 | |
| KIF17 | GDPD5 | KCNJ12 | |
| ESRRB | SCUBE1 | PIWIL3 | |
| NPY | SPON2 | C14orf180 | |

Fig. 5.2.1.
AR binding motif that is over-represented in regions adjacent to the top correlated segments (top panel). The consensus AR binding motif is shown as a reference (bottom panel).



## 5.3 Association of methylation signatures with genomic copy number alterations

Next, I extracted genome-wide copy number profiles from seven plasma samples subjected to both low passage whole genome sequencing (LP-WGS) and low passage whole genome bisulfite sequencing (LP-WGBS) and confirmed high degree of agreement between results from the same sample with and without bisulfite treatment (**Fig. 3.3.1.**). Using LP-WGBS from mCRPC plasma samples, I observed copy number alterations at a frequency consistent with previously described studies of mCRPC tissue or plasma [39,40] (for example, most commonly: 8q21-24 gain: prevalence $\geq$ 70%; Xq12 gain: prevalence $\geq$ 60%; 8p21 loss: prevalence $\geq$ 50%, **Fig. 5.3.1.**). I observed more copy number changes with increasing PC1 values, as an increasing tumour fraction improved copy number detection (**Fig. 5.3.2.**). Later, I suspected that methylation signatures may be biased by copy number alterations, and thus I then confirmed ct-MethSig or AR-MethSig were not located more frequently in regions of copy number alterations in our dataset (**Table. 5.3.1.**). To integrate genomic copy number data with specific methylation signatures, I evaluated the correlation of copy number status of every segment across the genome and PC1 values (Kruskal-Wallis test **Fig. 5.3.3**). Most notably, I identified a significant difference in PC3 value distributions between *AR* copy number gain and *AR* non-gain samples ($P$ = 0.018, Kruskal-Wallis test, **Fig. 5.3.4.**).

Fig. 5.3.1.
Prevalence of gain and loss events lined by chromosome position extracted from LP-WGBS on mCRPC plasma samples.

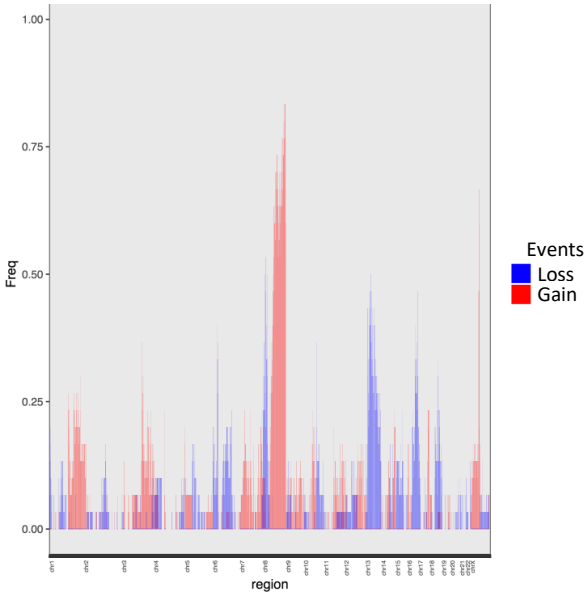


Fig. 5.3.2.
Analysis of copy number profiles on low-pass whole genome bisulfite sequencing. Matrix shows gains (red) and losses (blue) ordered by chromosomal position (columns) for individual patient samples (one per row) ordered by tumour purity. Bar chart on the left shows tumour fraction per sample. Bar chart on the right shows the number of gain (red) or loss (blue) events per sample

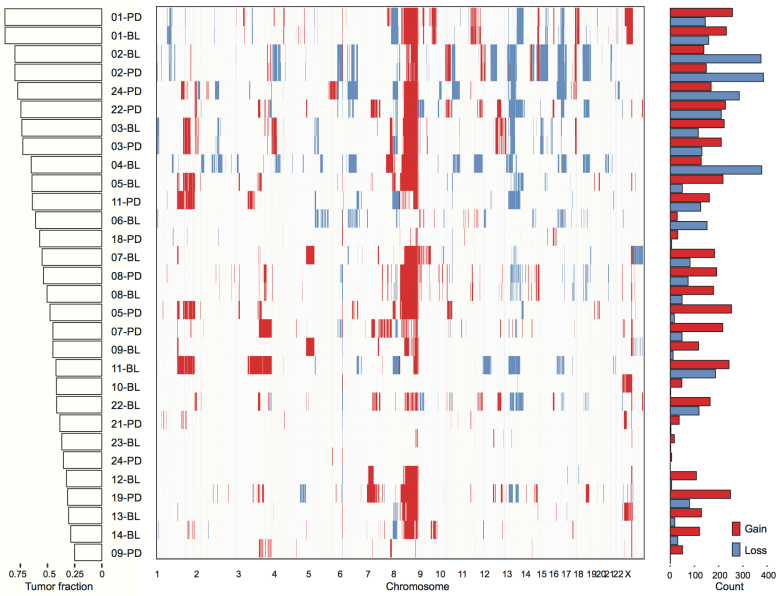



Fig. 5.3.3.
Manhattan plot showing the level of significance of the association between PC1 value distribution and copy number alterations ordered by chromosome position. The segment containing AR is highlighted as green dot (not significant, P = 0.18).

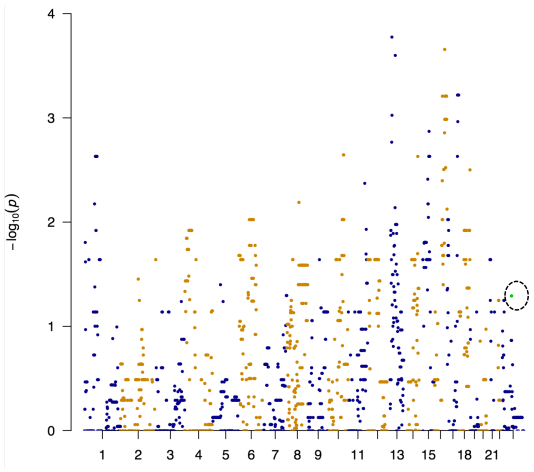


Fig. 5.3.4.
Manhattan plot showing the level of significance of the association between PC3 value distribution and copy number alterations ordered by chromosome position. The segment containing *AR* is highlighted as a green dot (*P* = 0.018, Kruskal-Wallis test).

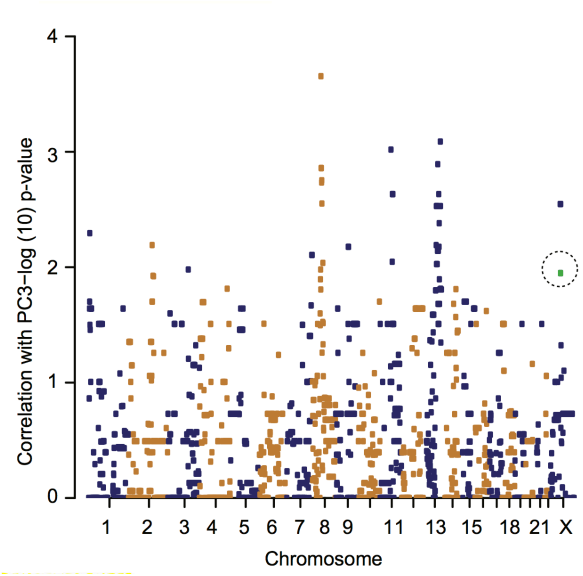


Table 5.3.1.
Contingency tables showing ct-MethSig and AR-MethSig segments in copy number aberrant regions.

| | CNA regions | non-CNA regions |
|---|---|---|
| ct-MethSig | 35 | 965 |
| All segments | 1031 | 236479 |
| | | |
| ar-MethSig | 0 | 1000 |
| All segments | 1031 | 236479 |

## 5.4 The AR-regulatory methylation signature may identify distinct clinical phenotypes

Based on the association of PC3 values and *AR* copy number status I confirmed that patient plasma and tissue samples with *AR* copy number gain had significantly lower AR-MethSig methylation ratio than *AR* copy number normal samples (Wilcoxon signed-rank test; **Fig. 5.4.1**). I also confirmed a high agreement for AR-MethSig extracted from high-coverage targeted NGS and LP-WGBS (95% limits of agreement: -0.136 to 0.076; **Fig. 5.4.2.**), and this introduces the opportunity to identify patient-specific signatures using LP-WGBS that is amenable to clinical implementability and scalability. I did not identify any hormone-sensitive cancers harbouring a low AR-MethSig median methylation ratio and nor did either of the two commonly studied AR-regulated prostate cancer cell lines (LNCaP and VCaP, **Fig. 5.1.2.**). Moreover, I was interested in evaluating the clinical relevance of AR-MethSig and as I had not observed a change over time in AR-MethSig median methylation ratio, I chose fixed time-points over the disease independent of the time of sampling: namely time from start of ADT to death. I observed that AR-MethSig low (AR-MethSig median methylation ratio < 0.6) cancers had poor clinical prognosis (HR = 8.18, 95% CI = 1.93–34.76, *P* = 0.0044; Mantel-Cox log-rank test; **Fig. 5.4.3**).

Fig. 5.4.1.
Methylation ratio of AR-MethSig segments of *AR* gain and non-gain groups

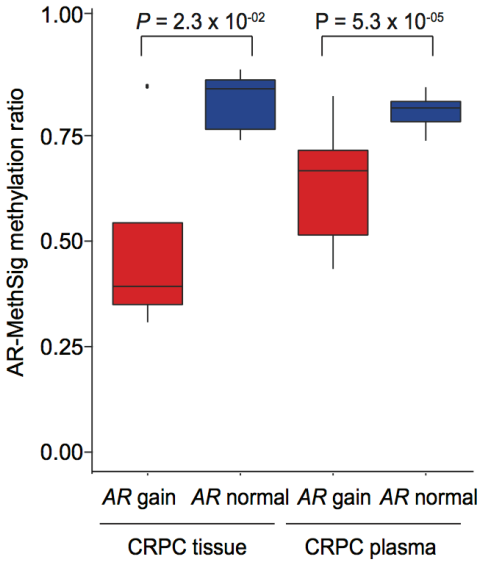
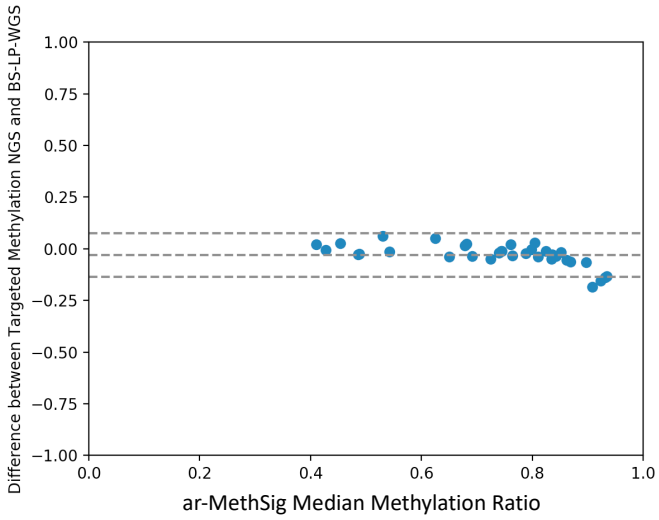
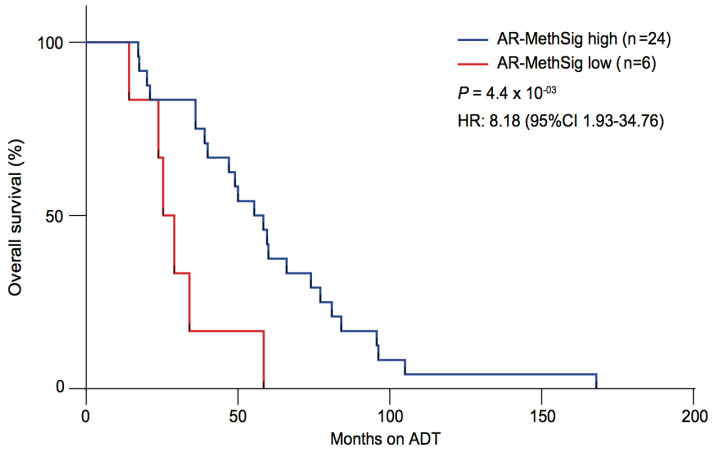



Fig. 5.4.2.
Bland-Altman plot showing agreement between targeted methylation NGS and LP-WGBS on AR-MethSig median methylation ratio




| | |
|---|---|
| **Bias** | **0.029** |
| **SD of bias** | **0.054** |
| **95% Limits of Agreement** | |
| **From** | **- 0.136** |
| **To** | **0.076** |

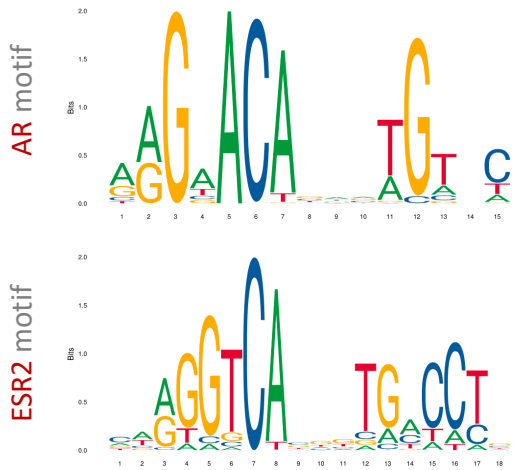
Fig. 5.4.3.
Overall survival analysis (start of ADT to death) for AR-MethSig low group versus AR-MethSig high group (Mantel-Cox log-rank test).



## 5.5 AR binding motif hypomethylation

Differentially methylated segments in AR-MethSig were observed in a subset of CRPC tumours, and this may be due to complex AR-regulatory mechanisms. I was intrigued to investigate the methylation status of AR-binding regions as I hypothesized that AR-binding may result in hypomethylation. I therefore extracted methylation ratio of all AR-binding sites described in JASPAR library – a well-documented database for transcription factor binding profiles. Meanwhile I also extracted estrogen receptor (ESR) binding sites as both transcriptional factors, AR and ESR, had very similar binding sequences (**Fig. 5.5.1.**).

Fig. 5.5.1.
Androgen receptor (AR) and estrogen receptor (ESR2) binding motif (from JASPAR library)




Across our pre-designed targeted capture panel (See Chapter 2), there were around 1,000 segments overlapping with an AR-binding motif. The result showed that AR-binding sites were hypomethylated in all prostate tissues (LNCaP, VCaP cell lines, and normal prostate tissues) as compared to the healthy volunteer

plasma samples which are primarily constituted by white blood cell DNA (**Fig. 5.5.2.**). In ESR binding regions, no hypomethylation events were observed in prostate-related tissues as compared with healthy volunteer plasma samples (**Fig. 5.5.2.**). In addition, to further confirm that the hypomethylation events were not primarily dominant in transcriptional factor binding sites, I looked into methylation ratios of MYC and p53 binding motifs, and the results indicated no prominent hypomethylation across different tissue types (**Fig. 5.5.3.**).

Fig. 5.5.2.
Methylation ratio distributions of AR and ESR2 binding motif across different tissue types

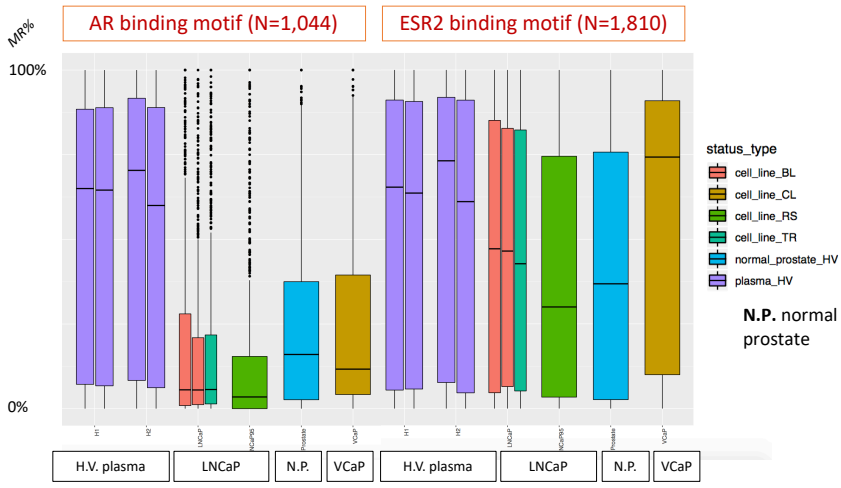
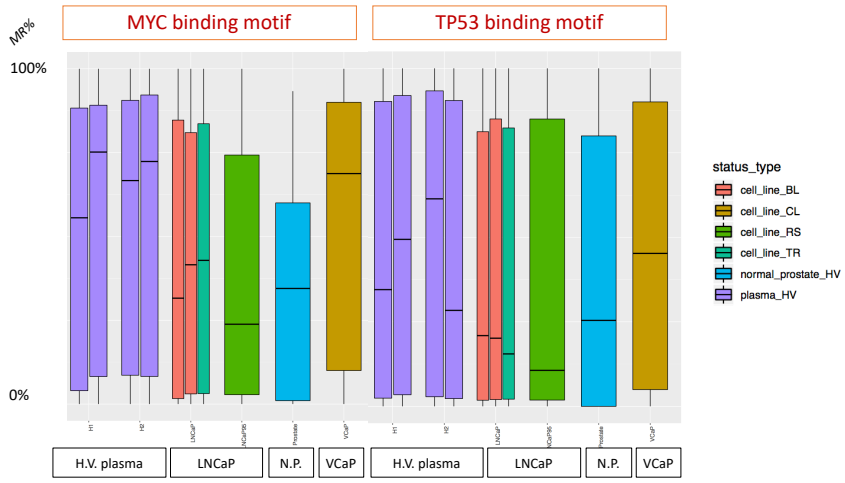


Fig. 5.5.3.
Methylation ratio distributions of MYC and TP53 binding motif across different tissue types




In conclusion, I observe hypomethylation events primarily in AR-binding motifs in prostate-related tissues, and more studies are required to elucidate the complex epigenetic regulatory mechanisms (See **5.6 Discussion**).

## 5.6 Discussion - Tumour subtyping based on DNA methylation signatures

### 5.6.1 Challenges in DNA methylation-based classification

DNA methylation has been used in other tumour types for classification such as central nervous system tumours which could be classified into more than 100 known tumour types by using DNA methylation-based machine learning classifier [126]. The molecularly-defined subtypes showed a high-level of standardisation and reduced substantially inter-observer and inter-institutional variability. However, to construct a DNA methylation-based subtyping system in prostate cancer based on plasma methylome could still be difficult. As the majority of methylation features extracted from plasma DNA are related to tumour DNA fraction, extracting methylation information specifically related to an individual's cancer could be challenging across a range of tumour fractions as seen in clinical practice and as exemplified in our cohort. Also, the lack of higher quality NGS-based sequencing data could also hinder validation of any identified methylation signatures. Higher coverage NGS on more tumours may address this challenge, with capture of sufficient tumour-specific reads even at low circulating fractions. To date, methylation NGS data on large mCRPC cohorts linked to clinical outcomes remains limited – international efforts have focused on obtaining genomic and transcriptomic data from tumour biopsies [41] [52] [15]. In the study of Beltran et al. selected methylation markers from CRPC patients were used to classify tumours with neuro-endocrine differentiation.

Further, it is also possible to adjust pan-genome methylation level according to tumour fraction using latent variable-based analysis, a collection of mathematical methods aiming to explain complex relations between several variables. Then the comparison between different samples become feasible. For example, differentially methylated cytosines that exhibit a statistically significant difference between two samples or two groups of samples can be identified [127] [128] [129]. Differentially variable CpGs (DVC) can also be identified and suggest distinct biological processes private to a sample group [123]. Other than statistical comparison between groups, unsupervised hierarchical clustering can also be applied, along with clinical features, to identify clinically-relevant tumour subtypes.

### 5.6.2 Biological relevance of AR-MethSig

In summary, I was able to identify AR-MethSig from the mCRPC plasma methylome that appears to represent a sub-group of cancers characterised by a more aggressive clinical course and enriched for *AR* copy number gain and hypomethylation at putative AR binding sites. A preliminary cell line study also indicated that the hypomethylation patterns across the regions overlapping with actual AR-binding sites described in JASPAR library. It still remains unclear how AR binding results in hypomethylation events. This finding could result from either part of the prostate organogenesis or AR-regulated transcriptional pathway leading to methyl group removal from 5-methylcytosine (5mC). More normal prostate tissues can help understand this, and treatment of anti-androgen agents on normal prostate epithelium cell lines (eg, PrEC) and prostate cancer cell lines

(eg, LNCaP and VCaP) can elucidate whether this phenomenon is reversible. Clinically, studies in more patients and healthy volunteers are required to validate our methylation sub-typing signatures and confirm response prediction.

DNA methylation was widely believed to be irreversible and only alleviated by DNA replication until year 2009 when scientists discovered the ten eleven translocation protein (TET). TET is responsible for DNA demethylation following successive oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) [130] [131]. Both 5caC or 5fC can then converted back to unmodified cytosine. Most notably, TET catalysing 5mC oxidisation is dependent on alpha-ketoglutarate, an important metabolite of citric acid cycle (or TCA cycle).

The TET protein has been shown to bind to AR and AR-coactivator proteins and the presence of TET2 binding sites and CpG hydroxymethylation events have been found on the AR regulated KLK3 gene[132]. Recently, Takayama et al. also showed that TET2 could be repressed by androgen in prostate cancer [133]. As the mechanism of AR-regulated demethylation has not been fully explained, further exploring the role of TET and androgen-AR signalling could potentially lead to novel biological findings.

# 6 Chapter 6. Future Directions

## 6.1 Conclusions of current study

The goal of my PhD study was to profile plasma methylome of mCRPC (**Chapter 2**), to identify the methylation signature(s) associated with tumour fraction (**Chapter 3** and **4**) and also characterise the methylation signatures specific to a clinically relevant methylation signature for tumour sub-typing (**Chapter 5**).

To summarise my study:

1. I started by concurrently analysing the plasma DNA methylome and genome from patients with metastatic prostate cancer with a wide range of circulating tumour fractions. These were integrated with the methylome from cell lines, healthy volunteer plasma DNA and prostate cancer tissues.
2. I split the plasma DNA methylome into segments and used principal component analysis (PCA) and identified a methylation component that highly correlated with genomically-determined tumour fraction obtained using a range of approaches.
3. I extracted the methylation ratio from several thousand regions that were highly correlated with the genomically-determined tumour fraction where the top 1000 correlated segments were named ct-MethSig.
4. The median methylation value of ct-MethSig as a score can be used to measure ctDNA fraction. I also confirmed that the ct-MethSig score

correlated with genomic and methylation-based assessments of tumour fraction in CSPC and CRPC tissues.

5. Ct-MethSig was characterized by hypermethylation of targets of the polycomb repressor complex 2 components (SUZ12, EED, and H3K27ME3).
6. Deconvolution of ct-MethSig identified circulating tissue-specific and cancer-specific methylation regions.
7. I showed that the ct-MethSig score can be used on low-passage whole genome bisulfite sequencing (LP-WGBS) that is a potentially cost-effective and clinically scalable approach.
8. I identified an orthogonal component (principal component three) which showed weak correlation with tumour fraction but still contributed 8% of global methylation variance.
9. Principal component three top correlated segments (AR-MethSig) revealed enrichment for androgen receptor binding sequences and where hypomethylation of these segments associated with *AR* copy number gain.
10. AR-MethSig, a methylation signature found in a subset of CRPC samples, can identify tumours with a more aggressive clinical course.

Albeit these promising findings, there were some limitations in this study and ways to improve future studies which I will discuss in the following sections.

## 6.2 Future directions and opportunities

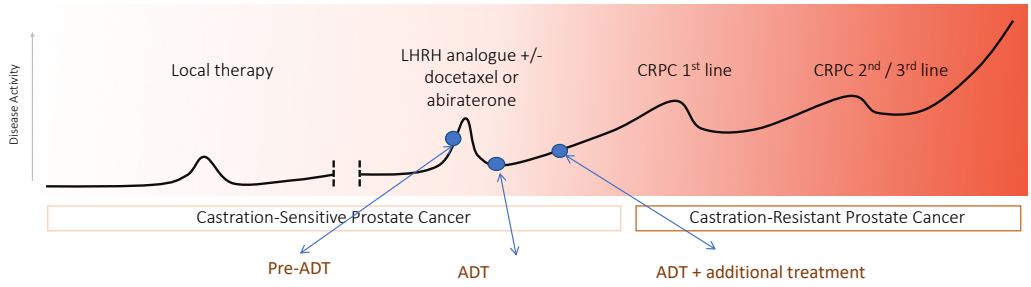
There are a lot of promises that cfDNA methylation analysis changing clinical practice and discovering novel cancer biology.

### 6.2.1 ct-MethSig in hormone-sensitive prostate cancer detection at relapse

The circulating tumour methylation signature or ct-MethSig can be used to accurately track tumour fraction changes, and this invention has the potential for detecting, screening, monitoring, risk classification for prostate cancer. The ct-MethSig contains both prostate-specific and prostate cancer-specific methylation signature. I plan to first test the signature in sequential plasma samples collected from high-risk, hormone-sensitive prostate cancer patients. Since 2015, the combination of ADT and additional treatment such as docetaxel, abiraterone acetate or enzalutamide has demonstrated survival advantages. For example, in CHAARTED study, the addition of chemotherapy-docetaxel could greatly improve the progression free survival for men with newly-diagnosed, metastatic hormone sensitive prostate cancer. However, most men eventually progressed with lethal metastatic hormone-resistant prostate cancer. There is an urgent need to improve the clinical management of these men. They may require additional tests for early detection of relapse, for better treatment selection or intensification and for interrogation of treatment resistance. ct-MethSig could be of great clinical value for early relapse detection. The common approach for plasma DNA analysis was to detect or measure the abundance of genomic alterations. However, this

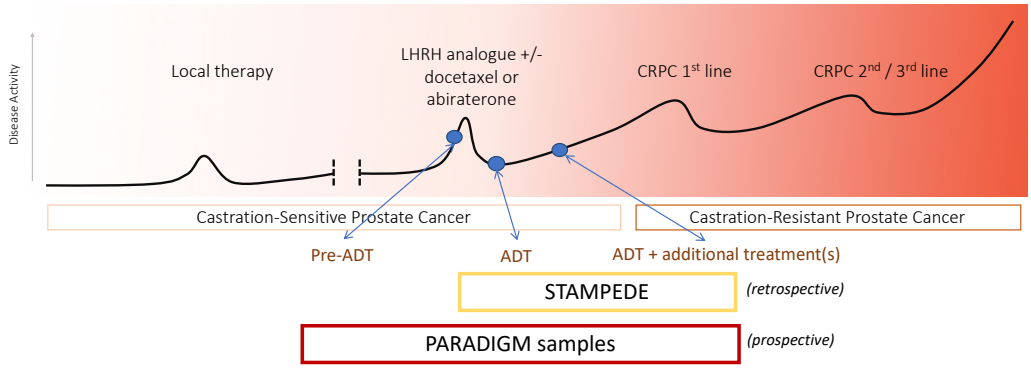
approach can be limited by the low prevalence of recurrent genomic changes, the relatively small number that are tumour specific and the low abundance in circulation of these aberrations that can overlap with other non-tumour aberrations, for example those resulting from clonal haematopoiesis. ct-MethSig takes thousands of CpG sites which are either prostate-specific or cancer-specific into account to measure tumour fraction, and can largely improve the detection sensitivity, especially at lower tumour abundance. I plan to perform whole genome bisulfite sequencing and obtain ct-MethSig methylation status on patients with newly diagnosed prostate cancer with or without metastatic diseases before the start of any treatment, on ADT, and on ADT ± additional treatments (Docetaxel, abiraterone, enzalutamide, or apalutamide; Figure 6.2.1.1.). I will use the pre-built ct-MethSig classification model (see **Chapter 4.**) and generate report which will include a) presence of circulating tumour DNA, b) circulating tumour DNA fraction and c) circulating prostate DNA fraction. I will also compare the plasma methylation analysis with PSA or testosterone level and other clinical parameters.

**Figure 6.2.1.1.**



Further I plan to perform feature engineering using whole genome bisulfite sequencing data from pre-ADT plasma samples (N > 30) and healthy volunteers (N > 30) and extract methylation status of newly-defined methylation features (e.g., CpG island with variable length, 100 base pairs long, fixed length segment), most of which are missed by the pre-designed target capture panel in my current study. Then I will optimise the ct-MethSig, including both prostate cancer-specific circulating and normal prostate-specific methylation signatures. The new ct-MethSig will be tested retrospectively and prospectively on clinical trials to study the clinical utility of MRD detection, especially for patients subject to primary treatment such as ADT ± additional treatments (Figure 6.2.1.2.).

**Figure 6.2.1.2.**




PARADIGM (Plasma Analysis for Response Assessment and to DIrect the manaGement of Metastatic prostate cancer), a prospective biomarker study, aims to collect plasma samples from newly diagnosed mCSPC patients before the start of ADT and sequentially along the treatment course. The major aim of this trial is

to understand whether the detection of ctDNA will be linked with worse clinical outcome (i.e., shorter time to CRPC or to death).

### 6.2.2 *AR*-regulated hypomethylation

In my preliminary study, I observed that androgen receptor binding motifs were pervasively hypomethylated in both normal and cancerous prostate tissues (see **Chapter 5.5**). I plan to validate the findings and elucidate *AR*-related epigenetic regulatory effects.

First, I will perform high coverage WGBS on hormone-sensitive prostate cancer and castration-resistance prostate cancer tissue and plasma samples. I will then extract known *AR* binding motifs and explore the methylation status of these regions. WGBS data from other tissue types such as white blood cell or breast tissue will be used as negative controls.

Chromatin accessibility is highly correlated with nucleosome positioning and it was demonstrated previously that pan-genome cfDNA coverage analysis allowed nucleosome positioning deconvolution [134]. Similarly, it was also shown in prior works that transcriptional factor binding can be used to infer nucleosome footprint, open chromatin status and gene expression [135,136]. Later, I will utilise high coverage WGBS from both HSPC and CRPC patients and investigate the genome-wide cfDNA coverage and fragment length across the *AR* binding motif. Ultimately, I aim to generate *AR*-regulated circulating prostate epithelium signature which can then be used to complement on ct-MethSig for superior cancer detection sensitivity.

### 6.2.3 Development of circulating methylation signature in other tumour types

The discovery that tumour fraction is the major determinant for pan-genome plasma methylome variance in metastatic prostate cancer has potential to be applied to other types of tumour and extract key methylation features for screening, relapse detection, monitoring, staging, risk stratification purposes. I plan to apply the same procedure (see **Chapter 2.4.**) to other cancer types such as bladder cancer. The methylation-based circulating tumour signature will be compared with clinical standards for screening and relapse detection and/or results from genomically-determined ctDNA assay.

## 6.3 Concluding remarks

Liquid biopsies allow repeated and clinically feasible collection of tumour material from metastatic patients. My study identifies methylation changes in 1000s of genomic segments that can be used to track circulating tumour DNA and potentially overcome some of the challenges inherent in genomic studies, including mutations due to aging without clear clinical significance [78], the paucity of common genomic events [39 40] and clonal hematopoiesis in older populations [137]. The plasma methylome could therefore represent an important source of additional information and currently remain underexplored in metastatic disease. In conclusion my study uses methylation features from plasma DNA to track circulating tumour fraction and identify sub-types of mCRPC with distinct biological mechanisms and differential clinical outcomes.

# 7 References


1 Fizazi, K. *et al.* Abiraterone plus Prednisone in Metastatic, Castration-Sensitive Prostate Cancer. *N Engl J Med* **377**, 352-360, doi:10.1056/NEJMoa1704174 (2017).

2 James, N. D. *et al.* Abiraterone for Prostate Cancer Not Previously Treated with Hormone Therapy. *N Engl J Med* **377**, 338-351, doi:10.1056/NEJMoa1702900 (2017).

3 Davis, I. D. *et al.* Enzalutamide with Standard First-Line Therapy in Metastatic Prostate Cancer. *N Engl J Med*, doi:10.1056/NEJMoa1903835 (2019).

4 Diaz, L. A., Jr. & Bardelli, A. Liquid biopsies: genotyping circulating tumor DNA. *J Clin Oncol* **32**, 579-586, doi:10.1200/JCO.2012.45.2011 (2014).

5 Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med* **20**, 548-554, doi:10.1038/nm.3519 (2014).

6 Janne, P. A. *et al.* AZD9291 in EGFR inhibitor-resistant non-small-cell lung cancer. *N Engl J Med* **372**, 1689-1699, doi:10.1056/NEJMoa1411817 (2015).

7 Ettinger, D. S. *et al.* Non-Small Cell Lung Cancer, Version 5.2017, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* **15**, 504-535, doi:10.6004/jnccn.2017.0050 (2017).

8 Wyatt, A. W. *et al.* Concordance of Circulating Tumor DNA and Matched Metastatic Tissue Biopsy in Prostate Cancer. *J Natl Cancer Inst* **109**, doi:10.1093/jnci/djx118 (2017).

9 Chakravarthy, A. *et al.* Pan-cancer deconvolution of tumour composition using DNA methylation. *Nat Commun* **9**, 3220, doi:10.1038/s41467-018-05570-1 (2018).

10 Benelli, M., Romagnoli, D. & Demichelis, F. Tumor purity quantification by clonal DNA methylation signatures. *Bioinformatics* **34**, 1642-1649, doi:10.1093/bioinformatics/bty011 (2018).

11 Altimari, A. *et al.* Diagnostic role of circulating free plasma DNA detection in patients with localized prostate cancer. *Am J Clin Pathol* **129**, 756-762, doi:10.1309/DBPX1MFNDDJBW1FL (2008).

12 Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci U S A* **112**, E5503-5512, doi:10.1073/pnas.1508736112 (2015).

13 Guo, S. *et al.* Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat Genet* **49**, 635-642, doi:10.1038/ng.3805 (2017).

14 Kang, S. *et al.* CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biol* **18**, 53, doi:10.1186/s13059-017-1191-5 (2017).

15 Quigley, D. A. *et al.* Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. *Cell* **175**, 889, doi:10.1016/j.cell.2018.10.019 (2018).

16 Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579-583, doi:10.1038/s41586-018-0703-0 (2018).

17 Zemmour, H. *et al.* Non-invasive detection of human cardiomyocyte death using methylation patterns of circulating DNA. *Nat Commun* **9**, 1443, doi:10.1038/s41467-018-03961-y (2018).

18 Lehmann-Werman, R. *et al.* Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc Natl Acad Sci U S A* **113**, E1826-1834, doi:10.1073/pnas.1519286113 (2016).

19 Moss, J. *et al.* Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat Commun* **9**, 5068, doi:10.1038/s41467-018-07466-6 (2018).

20 Ferlay, J. *et al.* Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer* **127**, 2893-2917, doi:10.1002/ijc.25516 (2010).

21 Jemal, A. *et al.* Global cancer statistics. *CA Cancer J Clin* **61**, 69-90, doi:10.3322/caac.20107 (2011).

22 Caverly, T. J. *et al.* Presentation of Benefits and Harms in US Cancer Screening and Prevention Guidelines: Systematic Review. *J Natl Cancer Inst* **108**, djv436, doi:10.1093/jnci/djv436 (2016).

23 Abraham, N. E., Mendhiratta, N. & Taneja, S. S. Patterns of repeat prostate biopsy in contemporary clinical practice. *J Urol* **193**, 1178-1184, doi:10.1016/j.juro.2014.10.084 (2015).

24 Kasivisvanathan, V. *et al.* MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis. *N Engl J Med* **378**, 1767-1777, doi:10.1056/NEJMoa1801993 (2018).

25 Tannock, I. F. *et al.* Docetaxel plus prednisone or mitoxantrone plus prednisone for advanced prostate cancer. *N Engl J Med* **351**, 1502-1512, doi:10.1056/NEJMoa040720 (2004).

26 Sweeney, C. J. *et al.* Chemohormonal Therapy in Metastatic Hormone-Sensitive Prostate Cancer. *N Engl J Med* **373**, 737-746, doi:10.1056/NEJMoa1503747 (2015).

27 James, N. D. *et al.* Addition of docetaxel, zoledronic acid, or both to first-line long-term hormone therapy in prostate cancer (STAMPEDE): survival results from an adaptive, multiarm, multistage, platform randomised controlled trial. *Lancet* **387**, 1163-1177, doi:10.1016/S0140-6736(15)01037-5 (2016).

28 Attard, G. *et al.* Selective inhibition of CYP17 with abiraterone acetate is highly active in the treatment of castration-resistant prostate cancer. *J Clin Oncol* **27**, 3742-3748, doi:10.1200/JCO.2008.20.0642 (2009).

29 Grist, E. & Attard, G. The development of abiraterone acetate for castration-resistant prostate cancer. *Urol Oncol* **33**, 289-294, doi:10.1016/j.urolonc.2015.03.021 (2015).

30 Yin, L. & Hu, Q. CYP17 inhibitors--abiraterone, C17,20-lyase inhibitors and multi-targeting agents. *Nat Rev Urol* **11**, 32-42, doi:10.1038/nrurol.2013.274 (2014).

31 de Bono, J. S. *et al.* Abiraterone and increased survival in metastatic prostate cancer. *N Engl J Med* **364**, 1995-2005, doi:10.1056/NEJMoa1014618 (2011).

32 Ryan, C. J. *et al.* Abiraterone in metastatic prostate cancer without previous chemotherapy. *N Engl J Med* **368**, 138-148, doi:10.1056/NEJMoa1209096 (2013).

33 Scher, H. I. *et al.* Increased survival with enzalutamide in prostate cancer after chemotherapy. *N Engl J Med* **367**, 1187-1197, doi:10.1056/NEJMoa1207506 (2012).

34 Beer, T. M. *et al.* Enzalutamide in metastatic prostate cancer before chemotherapy. *N Engl J Med* **371**, 424-433, doi:10.1056/NEJMoa1405095 (2014).

35 Clegg, N. J. *et al.* ARN-509: a novel antiandrogen for prostate cancer treatment. *Cancer Res* **72**, 1494-1503, doi:10.1158/0008-5472.CAN-11-3948 (2012).

36 Chi, K. N. *et al.* Apalutamide for Metastatic, Castration-Sensitive Prostate Cancer. *N Engl J Med*, doi:10.1056/NEJMoa1903307 (2019).

37 Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-677, doi:10.1016/j.cell.2013.03.021 (2013).

38 Fraser, M. *et al.* Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* **541**, 359-364, doi:10.1038/nature20788 (2017).

39 Robinson, D. *et al.* Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell* **162**, 454, doi:10.1016/j.cell.2015.06.053 (2015).

40 Annala, M. *et al.* Circulating Tumor DNA Genomics Correlate with Resistance to Abiraterone and Enzalutamide in Prostate Cancer. *Cancer Discov* **8**, 444-457, doi:10.1158/2159-8290.CD-17-0937 (2018).

41 Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215-1228, doi:10.1016/j.cell.2015.05.001 (2015).

42 Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* **128**, 683-692, doi:10.1016/j.cell.2007.01.029 (2007).

43 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).

44 Labbe, D. P. *et al.* Role of diet in prostate cancer: the epigenetic link. *Oncogene* **34**, 4683-4691, doi:10.1038/onc.2014.422 (2015).

45 Probst, A. V., Dunleavy, E. & Almouzni, G. Epigenetic inheritance during the cell cycle. *Nat Rev Mol Cell Biol* **10**, 192-206, doi:10.1038/nrm2640 (2009).

46 Hemberger, M., Dean, W. & Reik, W. Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. *Nat Rev Mol Cell Biol* **10**, 526-537, doi:10.1038/nrm2727 (2009).

47 Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev* **16**, 6-21, doi:10.1101/gad.947102 (2002).

48 Widschwendter, M. & Jones, P. A. DNA methylation and breast carcinogenesis. *Oncogene* **21**, 5462-5482, doi:10.1038/sj.onc.1205606 (2002).

49 Teschendorff, A. E. & Relton, C. L. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet* **19**, 129-147, doi:10.1038/nrg.2017.86 (2018).

50 Ballestar, E. & Esteller, M. Methyl-CpG-binding proteins in cancer: blaming the DNA methylation messenger. *Biochem Cell Biol* **83**, 374-384, doi:10.1139/o05-035 (2005).

51 Nakayama, T. *et al.* Epigenetic regulation of androgen receptor gene expression in human prostate cancers. *Lab Invest* **80**, 1789-1796 (2000).

52 Beltran, H. *et al.* Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat Med* **22**, 298-305, doi:10.1038/nm.4045 (2016).

53 Lee, W. H. *et al.* Cytidine methylation of regulatory sequences near the pi-class glutathione S-transferase gene accompanies human prostatic carcinogenesis. *Proc Natl Acad Sci U S A* **91**, 11733-11737 (1994).

54 Mahon, K. L. *et al.* Methylated Glutathione S-transferase 1 (mGSTP1) is a potential plasma free DNA epigenetic marker of prognosis and response to chemotherapy in castrate-resistant prostate cancer. *Br J Cancer* **111**, 1802-1809, doi:10.1038/bjc.2014.463 (2014).

55 Mahon, K. L. *et al.* Serum Free Methylated Glutathione S-transferase 1 DNA Levels, Survival, and Response to Docetaxel in Metastatic, Castration-resistant Prostate Cancer: Post Hoc Analyses of Data from a Phase 3 Trial. *Eur Urol*, doi:10.1016/j.eururo.2018.11.001 (2018).

56 Liu, L., Yoon, J. H., Dammann, R. & Pfeifer, G. P. Frequent hypermethylation of the RASSF1A gene in prostate cancer. *Oncogene* **21**, 6835-6840, doi:10.1038/sj.onc.1205814 (2002).

57 Aryee, M. J. *et al.* DNA methylation alterations exhibit intraindividual stability and interindividual heterogeneity in prostate cancer metastases. *Sci Transl Med* **5**, 169ra110, doi:10.1126/scitranslmed.3005211 (2013).

58 Brocks, D. *et al.* Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Rep* **8**, 798-806, doi:10.1016/j.celrep.2014.06.053 (2014).

59 Kim, J. H. *et al.* Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. *Genome Res* **21**, 1028-1041, doi:10.1101/gr.119347.110 (2011).

60 Borno, S. T. *et al.* Genome-wide DNA methylation events in TMPRSS2-ERG fusion-negative prostate cancers implicate an EZH2-dependent mechanism with miR-26a hypermethylation. *Cancer Discov* **2**, 1024-1035, doi:10.1158/2159-8290.CD-12-0041 (2012).

61 Cancer Genome Atlas Research, N. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011-1025, doi:10.1016/j.cell.2015.10.025 (2015).

62 Baena, E. *et al.* ETV1 directs androgen metabolism and confers aggressive prostate cancer in targeted mice and patients. *Genes Dev* **27**, 683-698, doi:10.1101/gad.211011.112 (2013).

63 Gebhard, C. *et al.* General transcription factor binding at CpG islands in normal cells correlates with resistance to de novo DNA methylation in

cancer cells. *Cancer Res* **70**, 1398-1407, doi:10.1158/0008-5472.CAN-09-3406 (2010).

64 Lin, P. C. *et al.* Epigenomic alterations in localized and advanced prostate cancer. *Neoplasia* **15**, 373-383, doi:10.1593/neo.122146 (2013).

65 Quigley, D. A. *et al.* Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. *Cell* **174**, 758-769 e759, doi:10.1016/j.cell.2018.06.039 (2018).

66 Lo, Y. M. *et al.* Presence of fetal DNA in maternal plasma and serum. *Lancet* **350**, 485-487, doi:10.1016/S0140-6736(97)02174-0 (1997).

67 Jiang, P. *et al.* Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc Natl Acad Sci U S A* **112**, E1317-1325, doi:10.1073/pnas.1500076112 (2015).

68 Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* **6**, 224ra224, doi:10.1126/scitranslmed.3007094 (2014).

69 Dawson, S. J. *et al.* Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med* **368**, 1199-1209, doi:10.1056/NEJMoa1213261 (2013).

70 Garcia-Murillas, I. *et al.* Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Sci Transl Med* **7**, 302ra133, doi:10.1126/scitranslmed.aab0021 (2015).

71 Abbosh, C. *et al.* Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **545**, 446-451, doi:10.1038/nature22364 (2017).

72 Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353-357, doi:10.1038/nature14347 (2015).

73 Carreira, S. *et al.* Tumor clone dynamics in lethal prostate cancer. *Sci Transl Med* **6**, 254ra125, doi:10.1126/scitranslmed.3009448 (2014).

74 Heitzer, E. *et al.* Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. *Genome Med* **5**, 30, doi:10.1186/gm434 (2013).

75 Adalsteinsson, V. A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun* **8**, 1324, doi:10.1038/s41467-017-00965-y (2017).

76 Choudhury, A. D. *et al.* Tumor fraction in cell-free DNA as a biomarker in prostate cancer. *JCI Insight* **3**, doi:10.1172/jci.insight.122109 (2018).

77 Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880-886, doi:10.1126/science.aaa6806 (2015).

78 Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911-917, doi:10.1126/science.aau3879 (2018).

79 Xu, R. H. *et al.* Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat Mater* **16**, 1155-1161, doi:10.1038/nmat4997 (2017).

80 Li, W. *et al.* CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA

methylation sequencing data. *Nucleic Acids Res* **46**, e89, doi:10.1093/nar/gky423 (2018).

81 Pinsky, P. F., Prorok, P. C. & Kramer, B. S. Prostate Cancer Screening - A Perspective on the Current State of the Evidence. *N Engl J Med* **376**, 1285-1289, doi:10.1056/NEJMsb1616281 (2017).

82 Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet* **50**, 928-936, doi:10.1038/s41588-018-0142-8 (2018).

83 Phallen, J. *et al.* Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med* **9**, doi:10.1126/scitranslmed.aan2415 (2017).

84 Tie, J. *et al.* Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. *Sci Transl Med* **8**, 346ra392, doi:10.1126/scitranslmed.aaf6219 (2016).

85 Le, D. T. *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409-413, doi:10.1126/science.aan6733 (2017).

86 Mateo, J. *et al.* DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer. *N Engl J Med* **373**, 1697-1708, doi:10.1056/NEJMoa1506859 (2015).

87 Heller, G. *et al.* Circulating Tumor Cell Number as a Response Measure of Prolonged Survival for Metastatic Castration-Resistant Prostate Cancer: A Comparison With Prostate-Specific Antigen Across Five Randomized Phase III Clinical Trials. *J Clin Oncol* **36**, 572-580, doi:10.1200/JCO.2017.75.2998 (2018).

88 Goodall, J. *et al.* Circulating Cell-Free DNA to Guide Prostate Cancer Treatment with PARP Inhibition. *Cancer Discov* **7**, 1006-1017, doi:10.1158/2159-8290.CD-17-0261 (2017).

89 Romanel, A. *et al.* Plasma AR and abiraterone-resistant prostate cancer. *Sci Transl Med* **7**, 312re310, doi:10.1126/scitranslmed.aac9511 (2015).

90 Alsop, K. *et al.* A community-based model of rapid autopsy in end-stage cancer patients. *Nat Biotechnol* **34**, 1010-1014, doi:10.1038/nbt.3674 (2016).

91 Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232, doi:10.1186/1471-2105-10-232 (2009).

92 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

93 Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res* **25**, 918-925, doi:10.1101/gr.176552.114 (2015).

94 Akalin, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* **13**, R87, doi:10.1186/gb-2012-13-10-r87 (2012).

95 Lê, S., Josse, J. & Husson, F. FactoMineR: An R Package for Multivariate Analysis. *2008* **25**, 18, doi:10.18637/jss.v025.i01 (2008).

96 Josse, J. & Husson, F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *2016* **70**, 31, doi:10.18637/jss.v070.i01 (2016).

97 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825-2830 (2011).

98 Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* **17**, 208, doi:10.1186/s13059-016-1066-1 (2016).

99 Prandi, D. *et al.* Unraveling the clonal hierarchy of somatic genomic aberrations. *Genome Biol* **15**, 439, doi:10.1186/s13059-014-0439-6 (2014).

100 Ha, G. *et al.* Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res* **22**, 1995-2007, doi:10.1101/gr.137570.112 (2012).

101 Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339, doi:10.1093/bioinformatics/bts378 (2012).

102 Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* **26**, 64-70, doi:10.1093/annonc/mdu479 (2015).

103 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219, doi:10.1038/nbt.2514 (2013).

104 Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* **46**, 912-918, doi:10.1038/ng.3036 (2014).

105 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-181, doi:10.1038/nmeth.1785 (2011).

106 Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-210 (2002).

107 Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat Commun* **6**, 8971, doi:10.1038/ncomms9971 (2015).

108 Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-16915, doi:10.1073/pnas.1009843107 (2010).

109 Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425, doi:10.1016/j.cels.2015.12.004 (2015).

110 Zambelli, F., Pesole, G. & Pavesi, G. PscanChIP: Finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments. *Nucleic Acids Res* **41**, W535-543, doi:10.1093/nar/gkt448 (2013).

111 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

112 Yu, G., Wang, L. G. & He, Q. Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382-2383, doi:10.1093/bioinformatics/btv145 (2015).

113 Lee, T. I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301-313, doi:10.1016/j.cell.2006.02.043 (2006).

114 Yu, J. *et al.* A polycomb repression signature in metastatic prostate cancer predicts cancer outcome. *Cancer Res* **67**, 10657-10663, doi:10.1158/0008-5472.CAN-07-2498 (2007).

115 Pidsley, R. *et al.* Enduring epigenetic landmarks define the cancer microenvironment. *Genome Res* **28**, 625-638, doi:10.1101/gr.229070.117 (2018).

116 Klus, P. *et al.* BarraCUDA - a fast short read sequence aligner using graphics processing units. *BMC Res Notes* **5**, 27, doi:10.1186/1756-0500-5-27 (2012).

117 Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790-2791, doi:10.1093/bioinformatics/btt468 (2013).

118 Liu, Y., Wirawan, A. & Schmidt, B. CUDASW++ 3.0: accelerating Smith-Waterman protein database search by coupling CPU and GPU SIMD instructions. *BMC Bioinformatics* **14**, 117, doi:10.1186/1471-2105-14-117 (2013).

119 Perez-Serrano, J., Sandes, E., Magalhaes Alves de Melo, A. C. & Ujaldon, M. DNA sequences alignment in multi-GPUs: acceleration and energy payoff. *BMC Bioinformatics* **19**, 421, doi:10.1186/s12859-018-2389-6 (2018).

120 Wilton, R. *et al.* Arioc: high-throughput read alignment with GPU-accelerated exploration of the seed-and-extend search space. *PeerJ* **3**, e808, doi:10.7717/peerj.808 (2015).

121 Wilton, R., Li, X., Feinberg, A. P. & Szalay, A. S. Arioc: GPU-accelerated alignment of short bisulfite-treated reads. *Bioinformatics* **34**, 2673-2675, doi:10.1093/bioinformatics/bty167 (2018).

122 Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572, doi:10.1093/bioinformatics/btr167 (2011).

123 Phipson, B. & Oshlack, A. DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biol* **15**, 465, doi:10.1186/s13059-014-0465-4 (2014).

124 Teschendorff, A. E. *et al.* DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nat Commun* **7**, 10478, doi:10.1038/ncomms10478 (2016).

125 Hyvarinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Netw* **13**, 411-430, doi:10.1016/s0893-6080(00)00026-5 (2000).

126 Capper, D. *et al.* DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469-474, doi:10.1038/nature26000 (2018).

127 Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-1369, doi:10.1093/bioinformatics/btu049 (2014).

128 Jaffe, A. E. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol* **41**, 200-209, doi:10.1093/ije/dyr238 (2012).

129 Hansen, K. D., Langmead, B. & Irizarry, R. A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* **13**, R83, doi:10.1186/gb-2012-13-10-r83 (2012).

130 He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303-1307, doi:10.1126/science.1210944 (2011).

131 Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300-1303, doi:10.1126/science.1210597 (2011).

132 Nickerson, M. L. *et al.* TET2 binds the androgen receptor and loss is associated with prostate cancer. *Oncogene* **36**, 2172-2183, doi:10.1038/onc.2016.376 (2017).

133 Takayama, K. *et al.* TET2 repression by androgen hormone regulates global hydroxymethylation status and prostate cancer progression. *Nat Commun* **6**, 8219, doi:10.1038/ncomms9219 (2015).

134 Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57-68, doi:10.1016/j.cell.2015.11.050 (2016).

135 Ulz, P. *et al.* Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat Genet* **48**, 1273-1278, doi:10.1038/ng.3648 (2016).

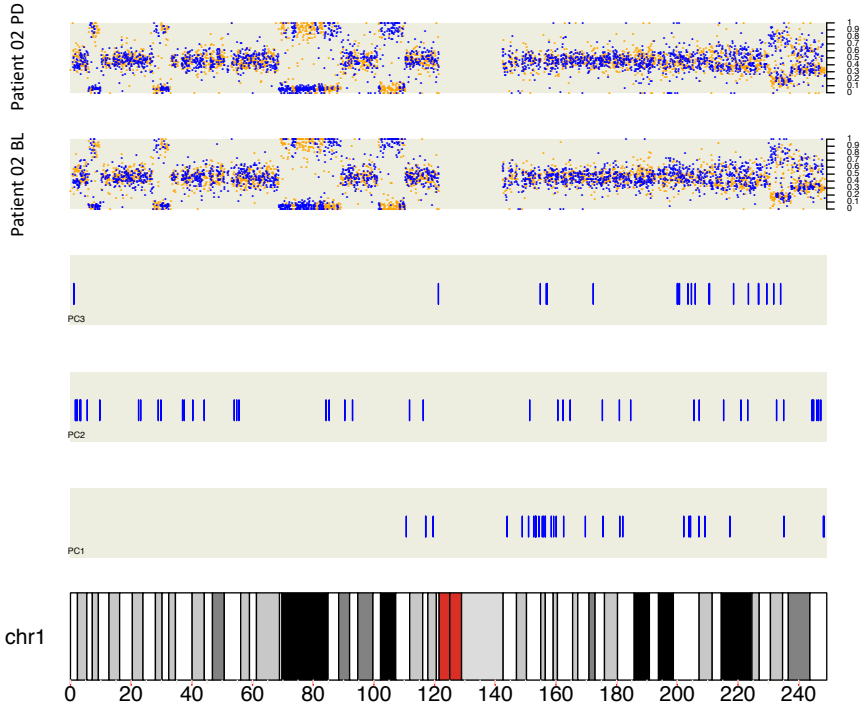
136 Ulz, P. *et al.* Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nat Commun* **10**, 4666, doi:10.1038/s41467-019-12714-4 (2019).

137 Mayrhofer, M. *et al.* Cell-free DNA profiling of metastatic prostate cancer reveals microsatellite instability, structural rearrangements and clonal hematopoiesis. *Genome Med* **10**, 85, doi:10.1186/s13073-018-0595-5 (2018).

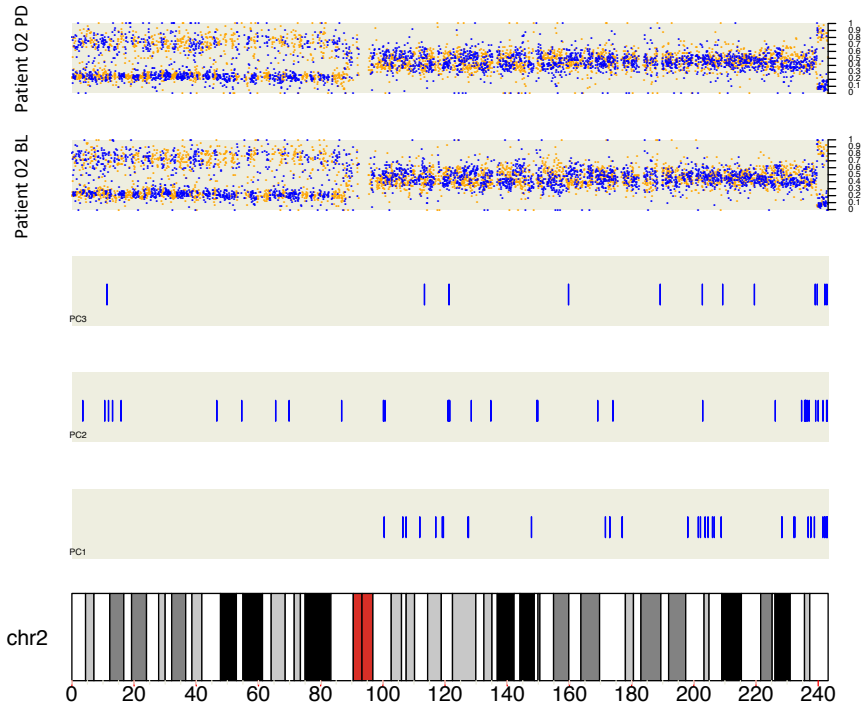
# 8 Supplementals.

- Supplementary Figure 3.7.4.
- Supplementary Table 5.2.3.

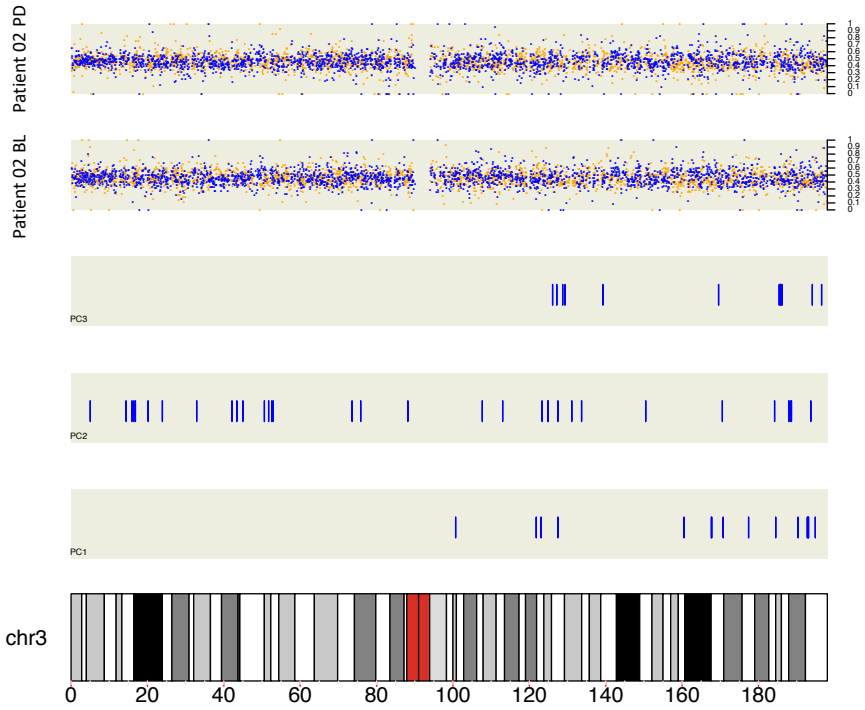
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



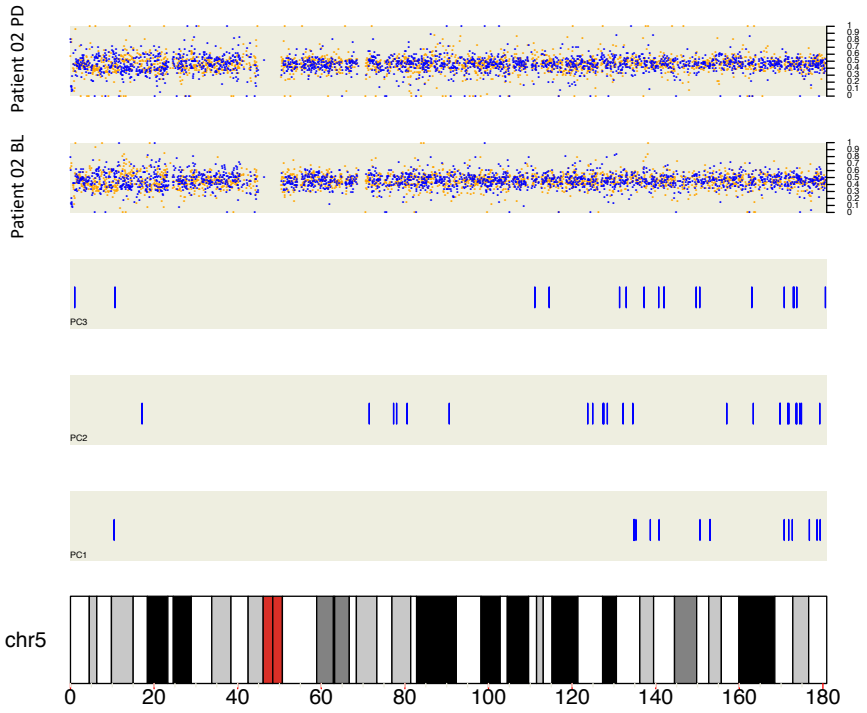
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



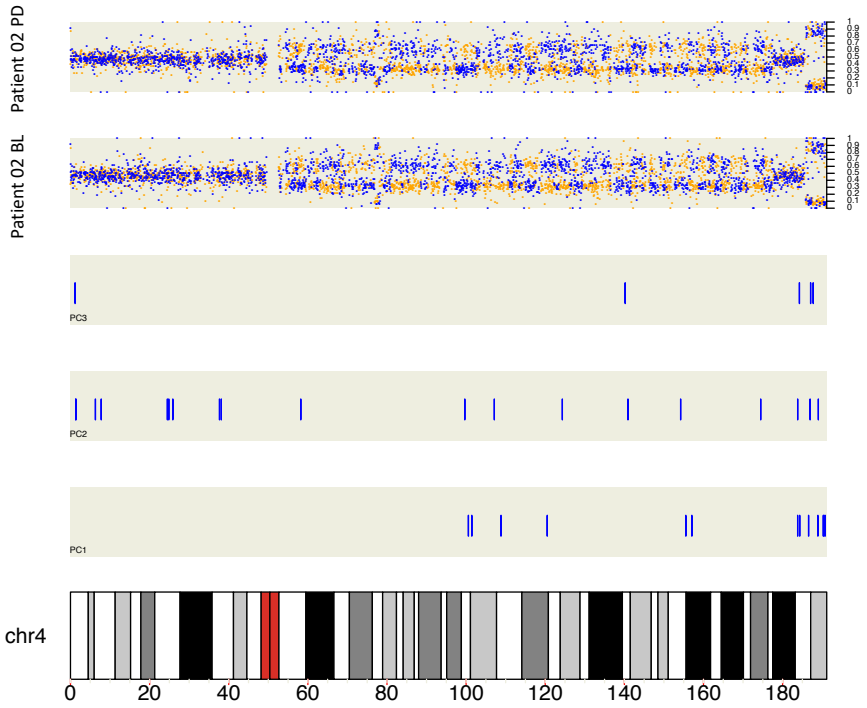
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



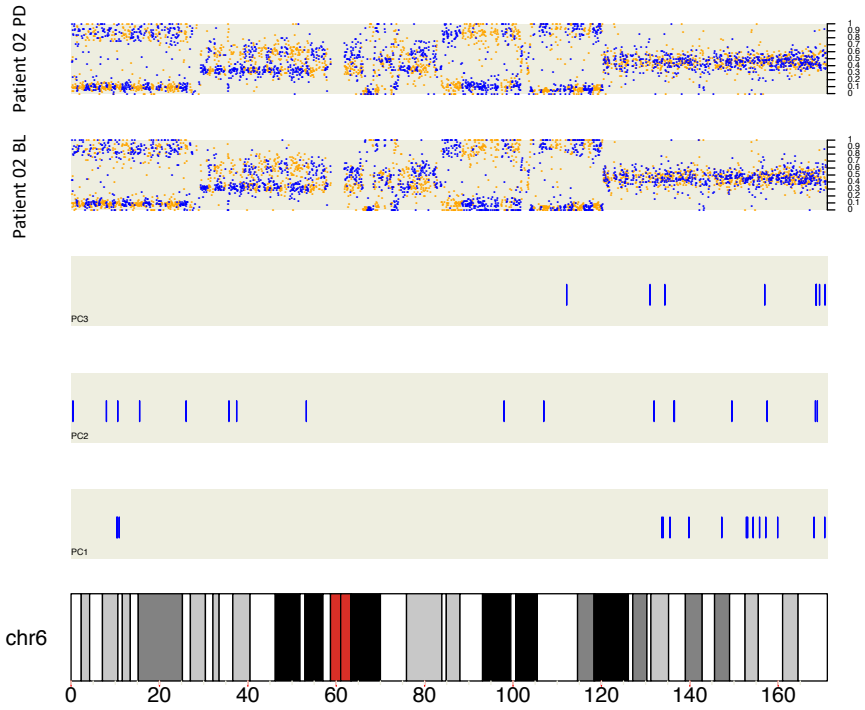
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



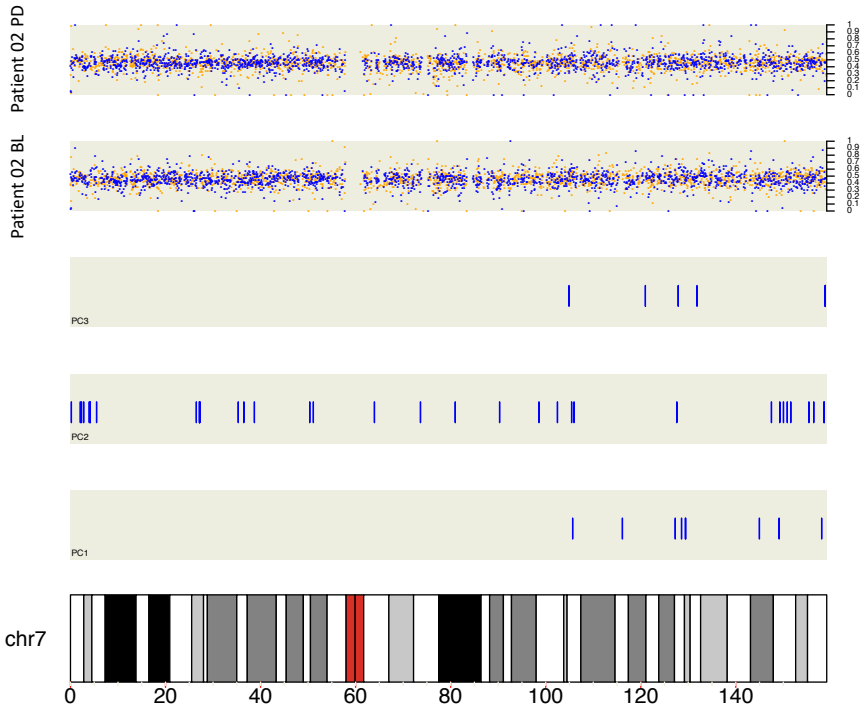
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



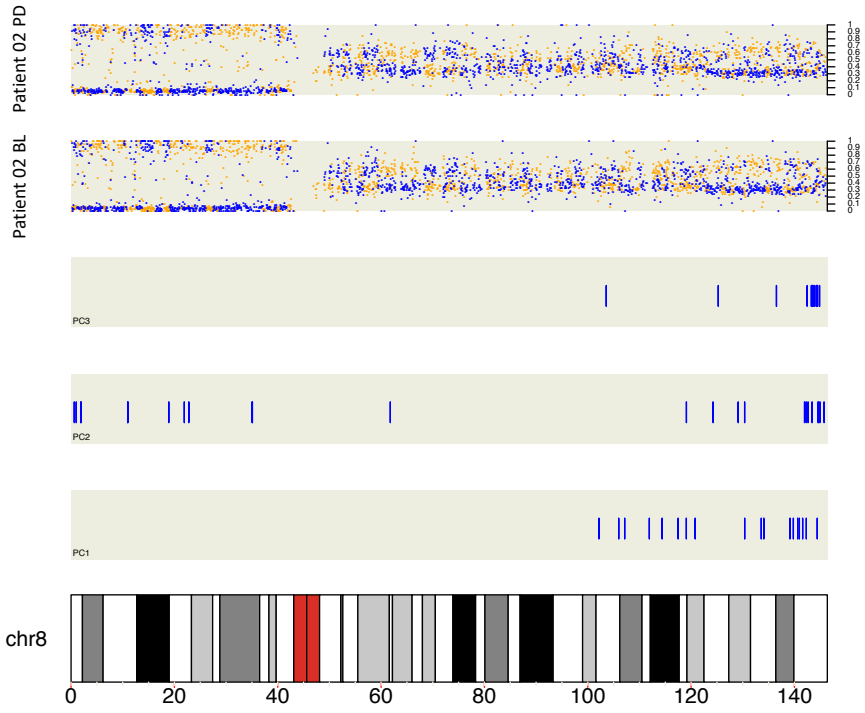
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



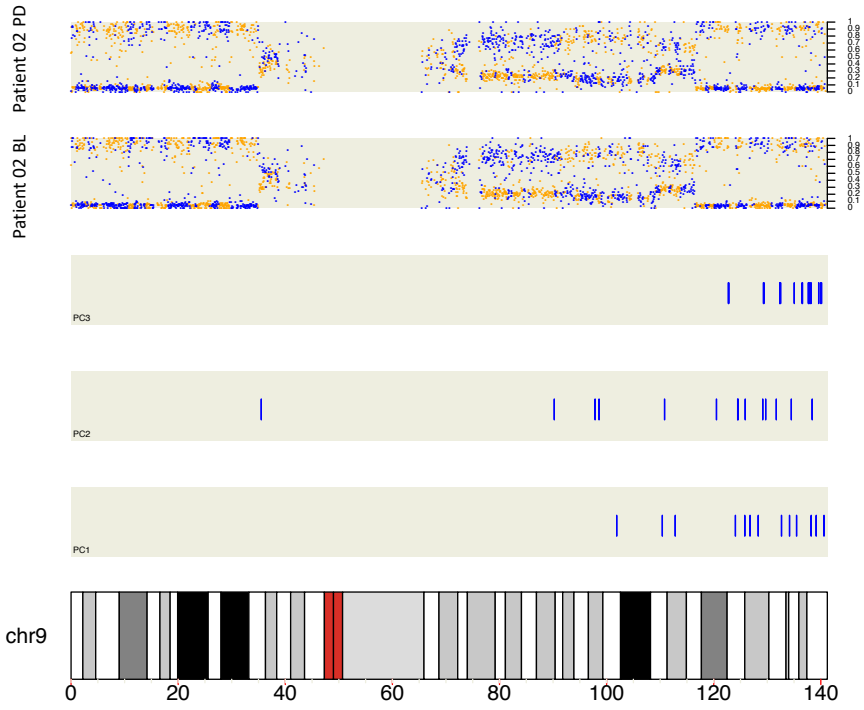
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



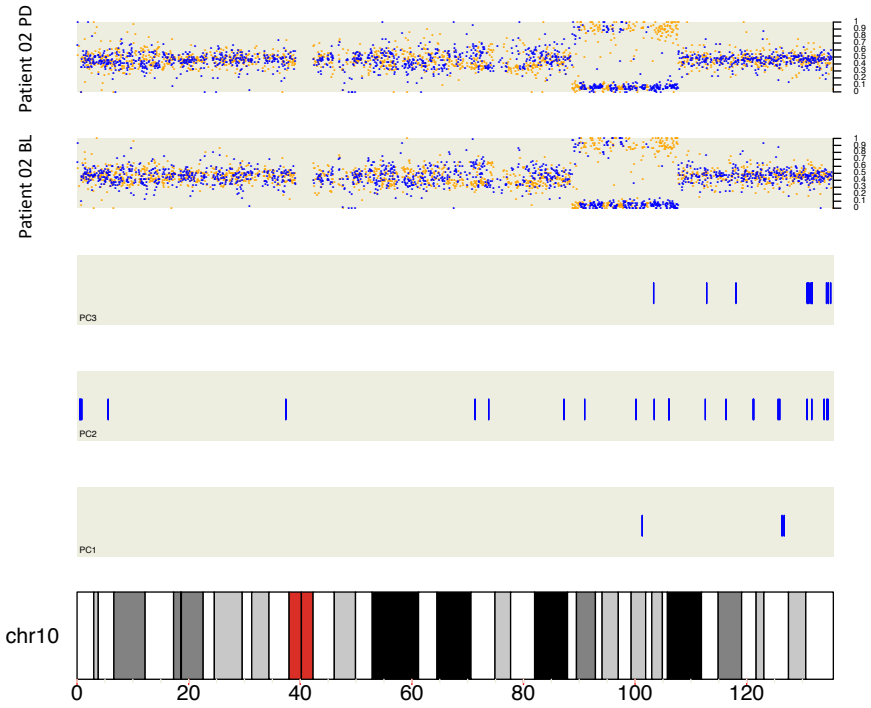
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



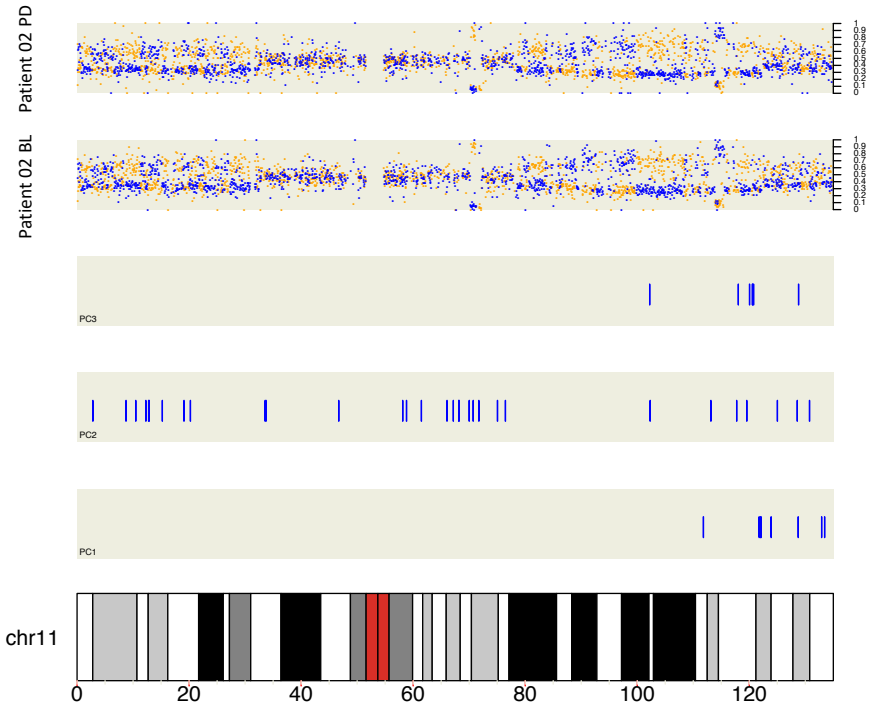
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



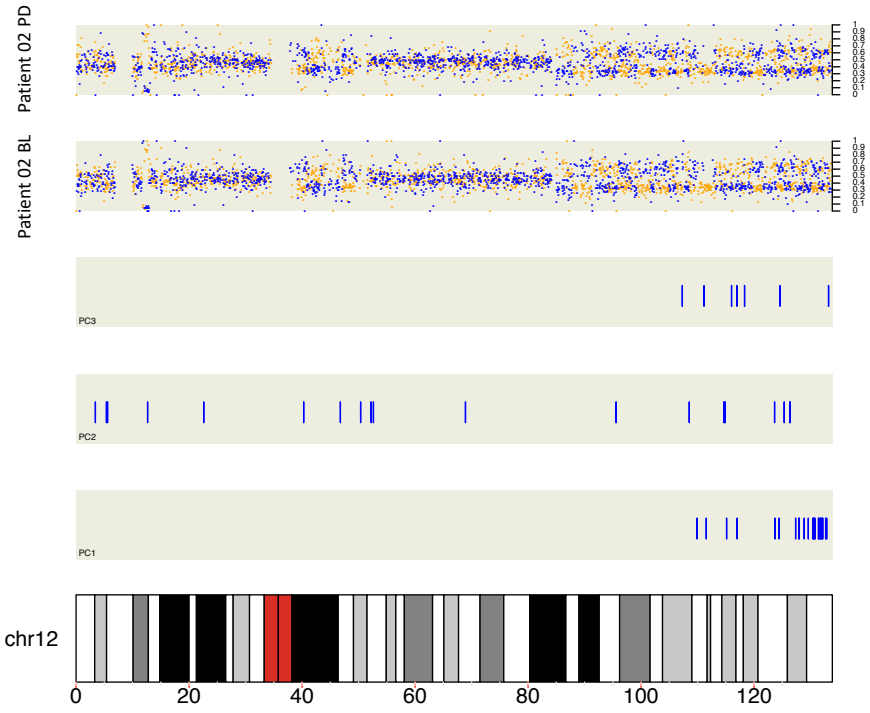
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



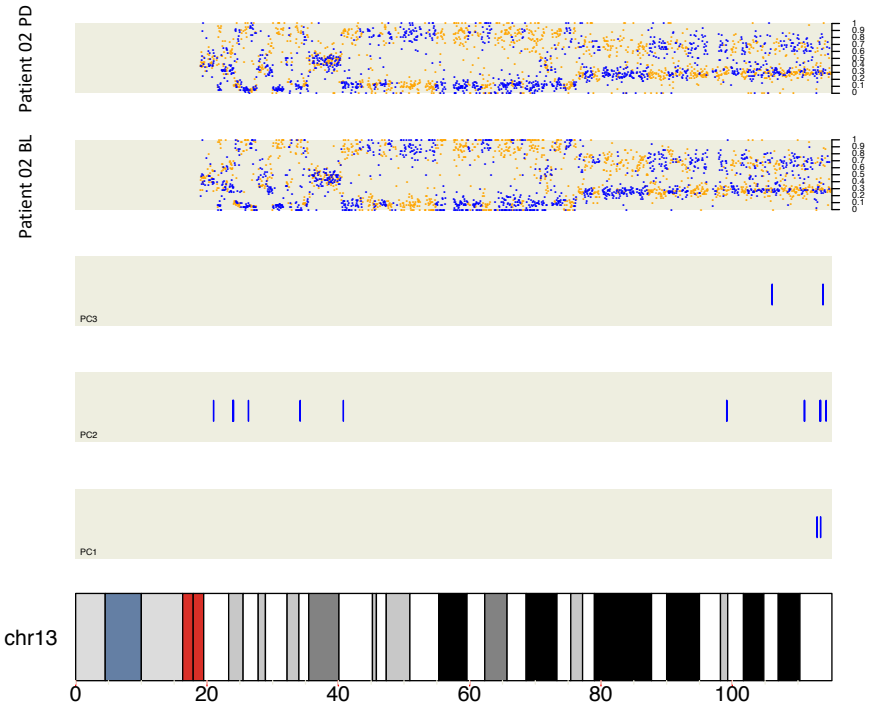
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



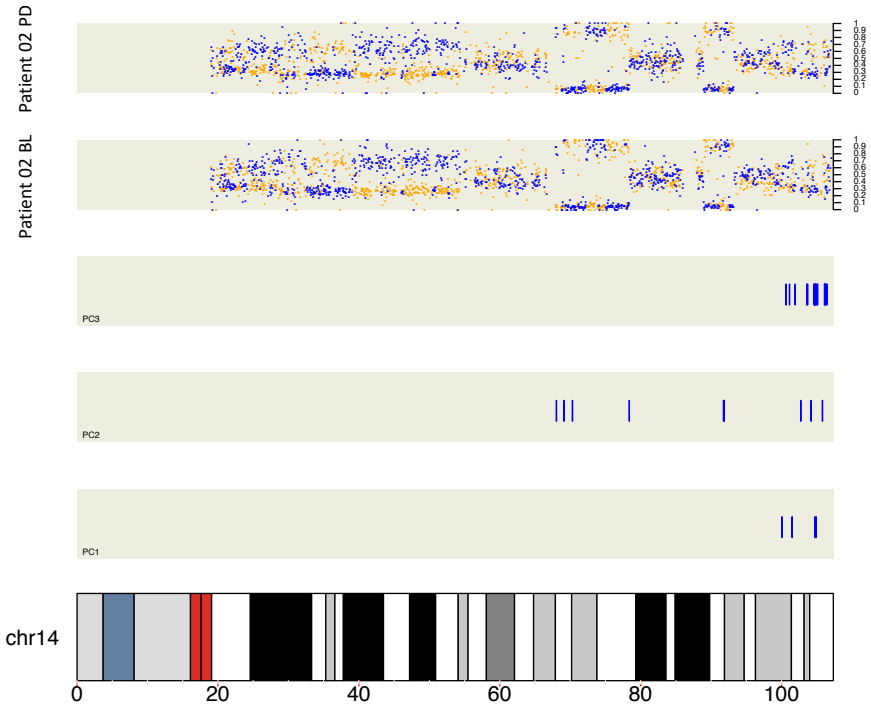
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



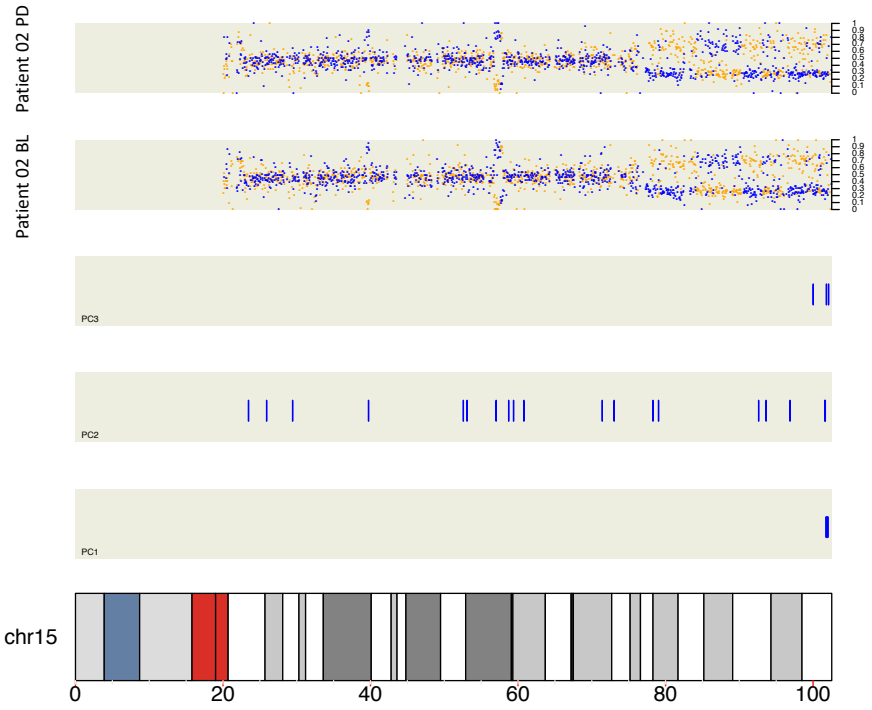
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



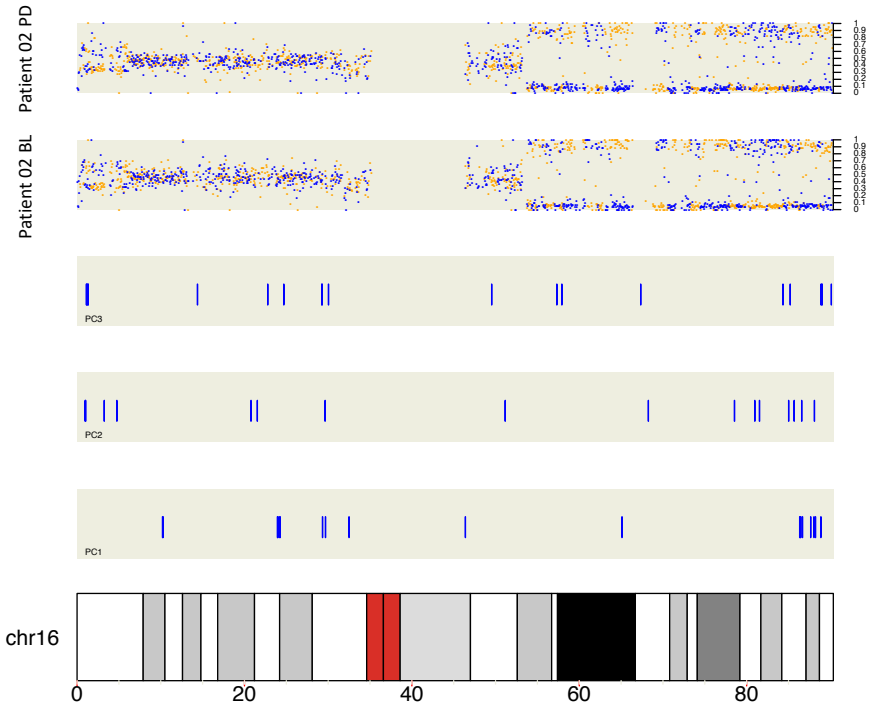
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



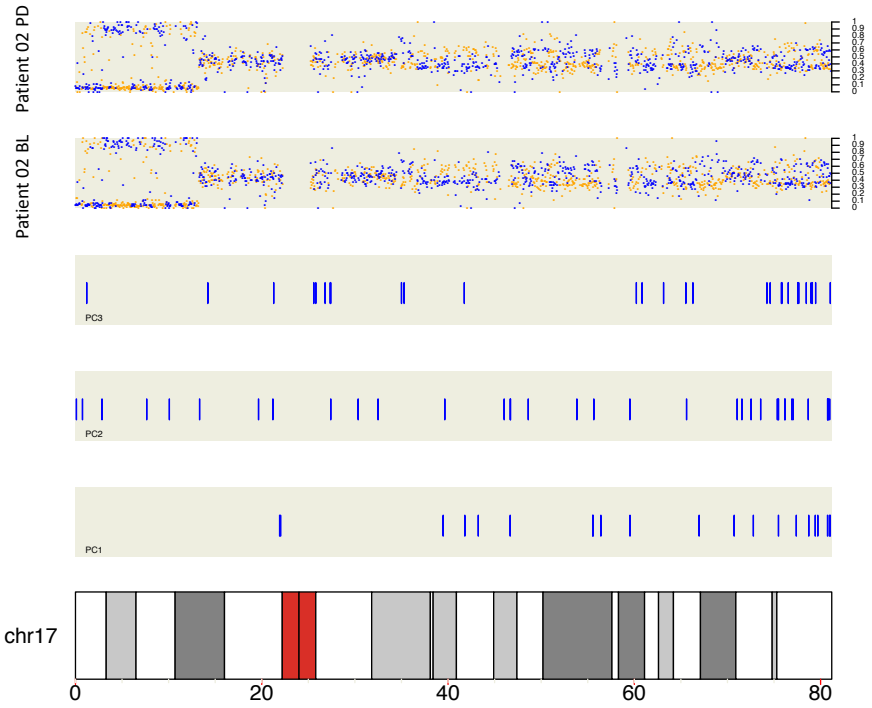
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



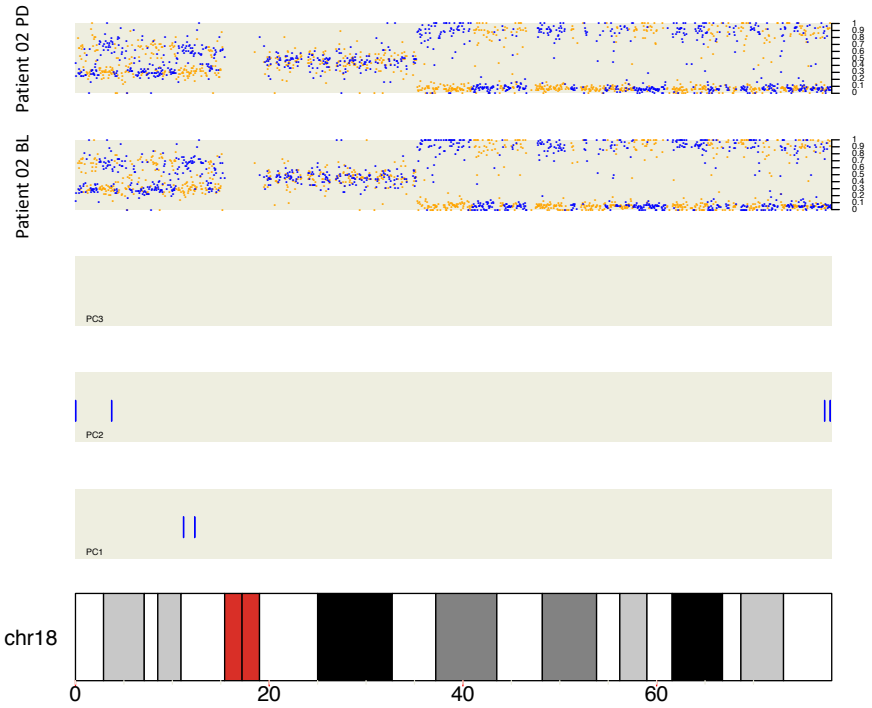
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



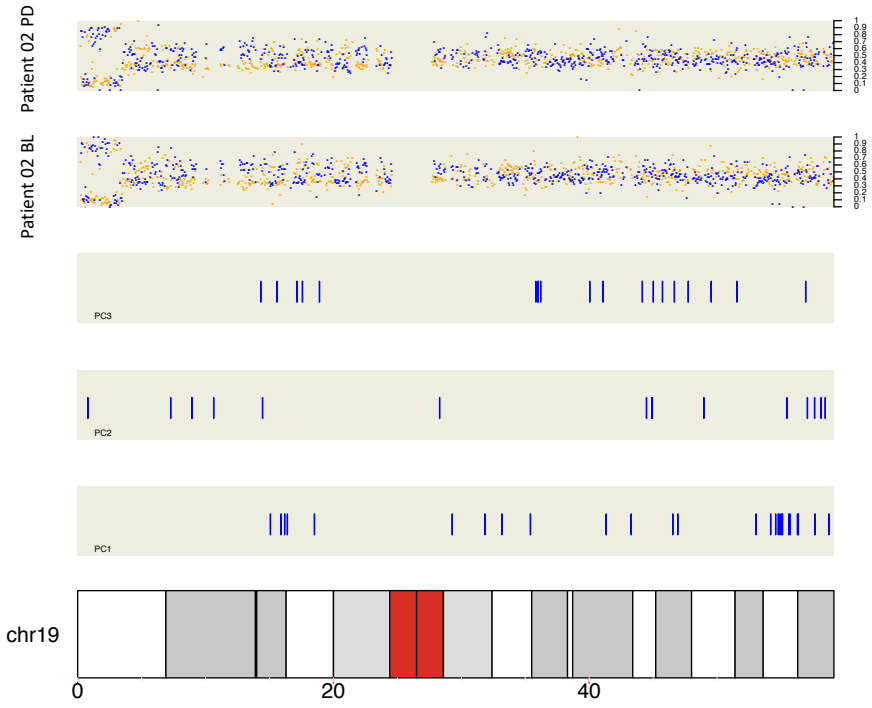
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



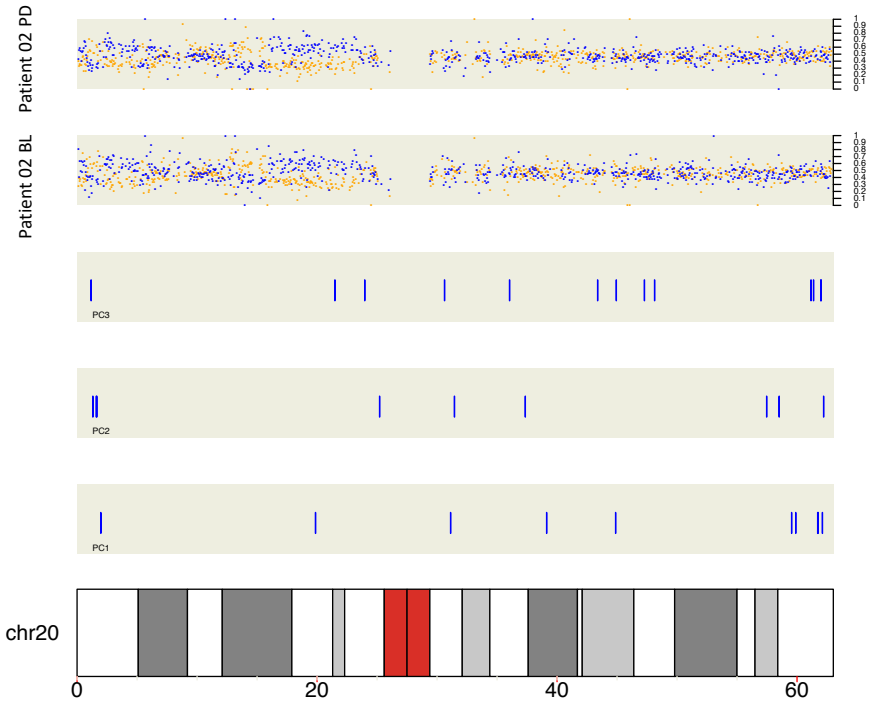
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



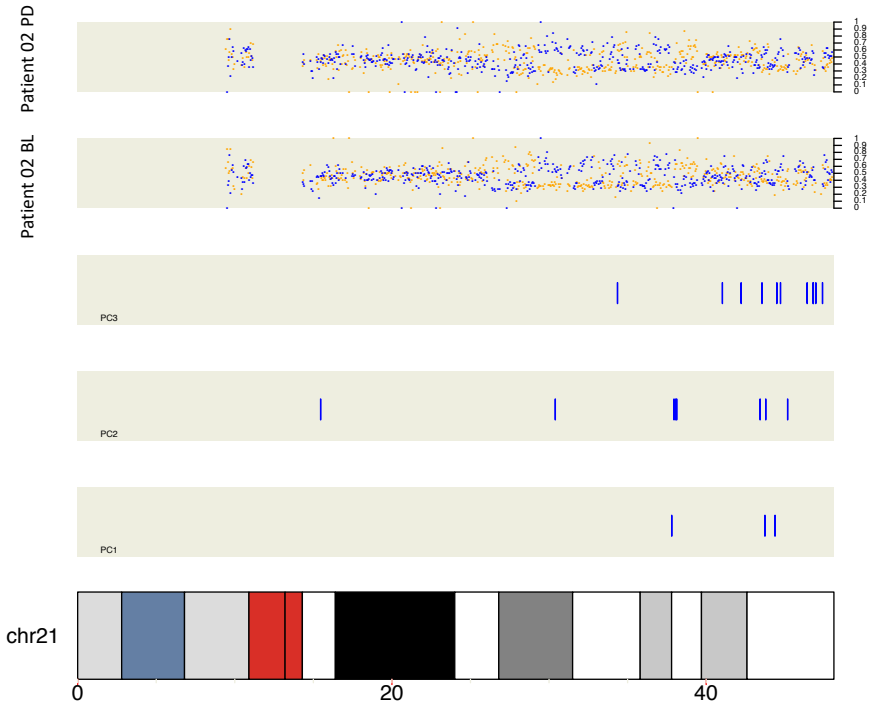
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



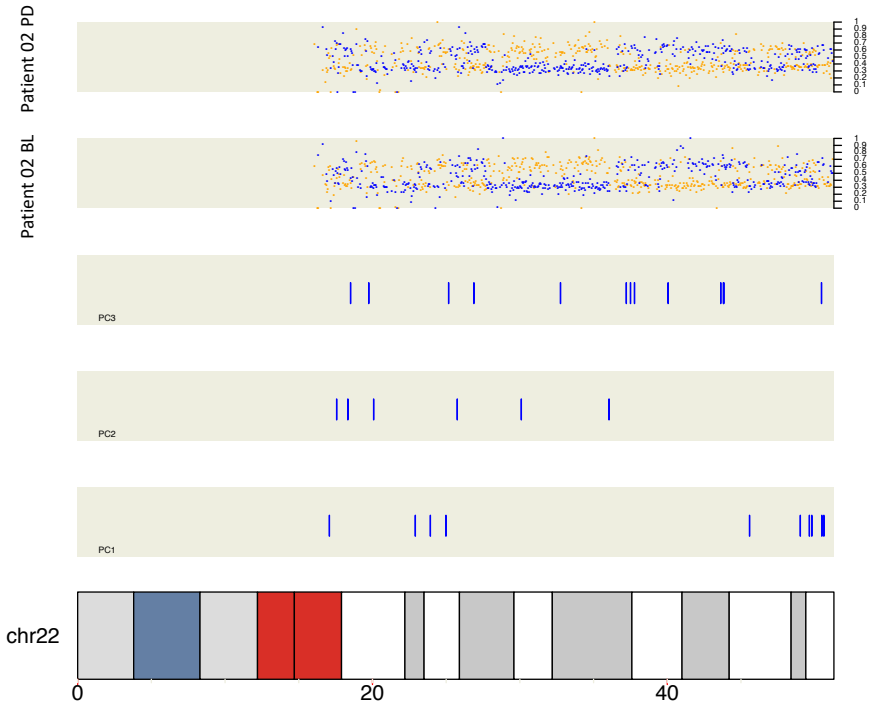
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



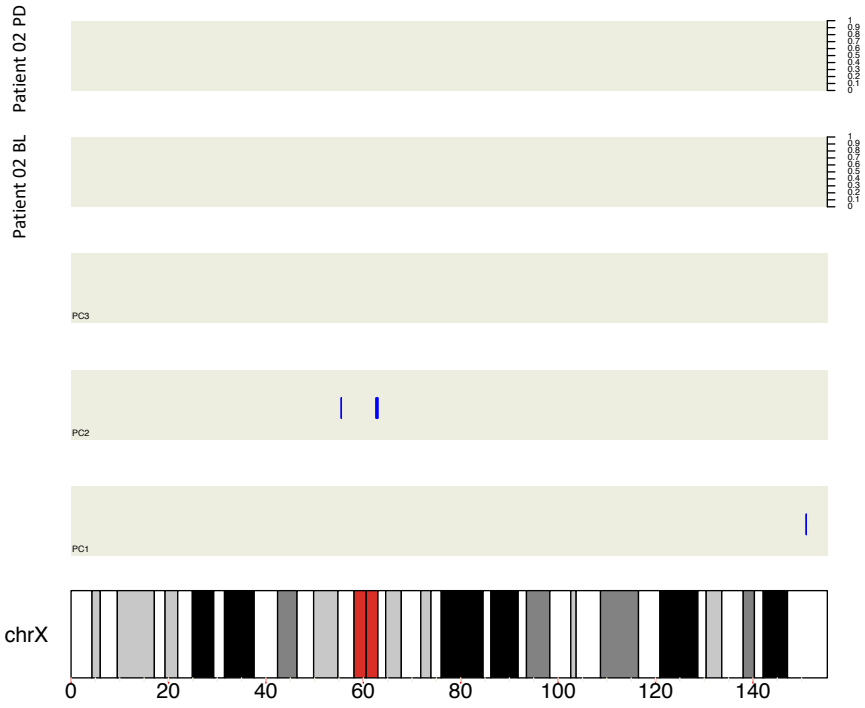
Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



Supplementary Figure 3.7.4.
Pan genomic B-allele frequency and PC1/2/3 distribution of patient 02 (trial ID: V5322)



Supplementary Table. 5.2.3. (1/10)
Functional enrichment of principal component three top 1000 segments

| CHR | START of ar-MethSig | END of ar-MethSig | REG_STRAN | ABS_SITE_START | ABS_SITE_END | REL_SITE_ST | REL_SITE_EN | SITE_STRAN | SCORE | SITE |
|---|---|---|---|---|---|---|---|---|---|---|
| chr14 | 105052526 | 105052675 | + | 105052565 | 105052581 | -36 | -21 | + | 0.876674 | GGGAACAGGATTTACCT |
| chr14 | 105052476 | 105052625 | + | 105052565 | 105052581 | 14 | 29 | + | 0.876674 | GGGAACAGGATTTACCT |
| chr14 | 105052426 | 105052575 | + | 105052565 | 105052581 | 64 | 79 | + | 0.876674 | GGGAACAGGATTTACCT |
| chr1 | 861676 | 861825 | + | 861753 | 861769 | 2 | 17 | - | 0.872295 | GGGGTCACCGTGTTCTC |
| chr4 | 7252376 | 7252525 | + | 7252514 | 7252530 | 63 | 78 | - | 0.871746 | AAGAACAGTGGGTTCCC |
| chr15 | 41219376 | 41219525 | + | 41219494 | 41219510 | 43 | 58 | + | 0.866725 | GGGTAGAGAGTGTTCCA |
| chr2 | 113379876 | 113380025 | + | 113379977 | 113379993 | 26 | 41 | + | 0.854291 | AGCCACACGGTGTGCCT |
| chr1 | 48360376 | 48360525 | + | 48360443 | 48360459 | -8 | 7 | + | 0.852908 | GTGAACATGACGTACCA |
| chr1 | 48360326 | 48360475 | + | 48360443 | 48360459 | 42 | 57 | + | 0.852908 | GTGAACATGACGTACCA |
| chr1 | 200707026 | 200707175 | + | 200707112 | 200707128 | 11 | 26 | + | 0.852876 | AAGAACAGGGTGCTCCG |
| chr1 | 200706976 | 200707125 | + | 200707112 | 200707128 | 61 | 76 | + | 0.852876 | AAGAACAGGGTGCTCCG |
| chr14 | 106320426 | 106320575 | + | 106320481 | 106320497 | -20 | -5 | + | 0.845683 | AGGGACACGTTGTACAG |
| chr2 | 121279826 | 121279975 | + | 121279826 | 121279842 | -75 | -60 | + | 0.840785 | GGGAATGTGCTGTACCC |
| chr2 | 121279776 | 121279925 | + | 121279826 | 121279842 | -25 | -10 | + | 0.840785 | GGGAATGTGCTGTACCC |
| chr2 | 121279726 | 121279875 | + | 121279826 | 121279842 | 25 | 40 | + | 0.840785 | GGGAATGTGCTGTACCC |
| chr15 | 60919376 | 60919525 | + | 60919386 | 60919402 | -65 | -50 | + | 0.832409 | GGGTACACAGTGATCAG |
| chr15 | 60919326 | 60919475 | + | 60919386 | 60919402 | -15 | 0 | + | 0.832409 | GGGTACACAGTGATCAG |
| chr1 | 15655876 | 15656025 | + | 15655958 | 15655974 | 7 | 22 | + | 0.831211 | AGGCACGCGCTGTCCTG |
| chr1 | 15655826 | 15655975 | + | 15655958 | 15655974 | 57 | 72 | + | 0.831211 | AGGCACGCGCTGTCCTG |
| chr5 | 142005126 | 142005275 | + | 142005141 | 142005157 | -60 | -45 | - | 0.830344 | TGGTACACCACGGTCCA |
| chr14 | 105044876 | 105045025 | + | 105044899 | 105044915 | -52 | -37 | + | 0.822933 | GAGCGCACTGTGTGCCA |
| chr14 | 105044826 | 105044975 | + | 105044899 | 105044915 | -2 | 13 | + | 0.822933 | GAGCGCACTGTGTGCCA |
| chr3 | 169540226 | 169540375 | + | 169540253 | 169540269 | -48 | -33 | - | 0.821421 | AGGAGCTCTGTGTTCTC |
| chr3 | 169540176 | 169540325 | + | 169540253 | 169540269 | 2 | 17 | - | 0.821421 | AGGAGCTCTGTGTTCTC |
| chr7 | 2959076 | 2959225 | + | 2959089 | 2959105 | -62 | -47 | + | 0.820473 | TGGTACATGGTGTCACG |
| chr7 | 2959026 | 2959175 | + | 2959089 | 2959105 | -12 | 3 | + | 0.820473 | TGGTACATGGTGTCACG |
| chr5 | 73969126 | 73969275 | + | 73969228 | 73969244 | 27 | 42 | - | 0.817906 | TGCCACAGAATGTTCCC |
| chr14 | 103691276 | 103691425 | + | 103691303 | 103691319 | -48 | -33 | - | 0.817436 | TGGTCCACTGCGTCCCT |
| chr14 | 103691226 | 103691375 | + | 103691303 | 103691319 | 2 | 17 | - | 0.817436 | TGGTCCACTGCGTCCCT |
| chr16 | 4673826 | 4673975 | + | 4673954 | 4673970 | 53 | 68 | - | 0.814584 | GGGAACACCTTGCACAC |
| chr9 | 136566976 | 136567125 | + | 136567043 | 136567059 | -8 | 7 | - | 0.814301 | CAGCACAGCATGAACCC |
| chr9 | 136566926 | 136567075 | + | 136567043 | 136567059 | 42 | 57 | - | 0.814301 | CAGCACAGCATGAACCC |
| chr3 | 64338476 | 64338625 | + | 64338481 | 64338497 | -70 | -55 | - | 0.813636 | AAGAACAGTATGTCTCT |
| chr3 | 64338426 | 64338575 | + | 64338481 | 64338497 | -20 | -5 | - | 0.813636 | AAGAACAGTATGTCTCT |
| chr17 | 60828076 | 60828225 | + | 60828091 | 60828107 | -60 | -45 | - | 0.810002 | CATCACACCCTGTTCCA |
| chr7 | 44279576 | 44279725 | + | 44279698 | 44279714 | 47 | 62 | + | 0.803908 | AGGTACAGCTTGTCCAG |
| chr4 | 140201576 | 140201725 | + | 140201587 | 140201603 | -64 | -49 | - | 0.803644 | TGGCTCATTCTGTACCA |
| chr4 | 140201526 | 140201675 | + | 140201587 | 140201603 | -14 | 1 | - | 0.803644 | TGGCTCATTCTGTACCA |
| chr10 | 3797376 | 3797525 | + | 3797383 | 3797399 | -68 | -53 | + | 0.802711 | GGGAACACGTCATGCTC |
| chr10 | 3797326 | 3797475 | + | 3797383 | 3797399 | -18 | -3 | + | 0.802711 | GGGAACACGTCATGCTC |
| chr4 | 26493326 | 26493475 | + | 26493377 | 26493393 | -24 | -9 | + | 0.802679 | GGGTCCTGGGTGTTCCA |
| chr5 | 170743776 | 170743925 | + | 170743888 | 170743904 | 37 | 52 | + | 0.79747 | GGGAAGGCGGTTTTCCT |
| chr7 | 65970026 | 65970175 | + | 65970143 | 65970159 | 42 | 57 | + | 0.797394 | CAGACCAGGTTGTACTC |
| chr8 | 125249776 | 125249925 | + | 125249789 | 125249805 | -62 | -47 | + | 0.796722 | GAGGACAGAGCTTTCCC |
| chr14 | 104623526 | 104623675 | + | 104623620 | 104623636 | 19 | 34 | + | 0.796695 | GGGCAATTAGTGTTCCT |
| chr16 | 2863726 | 2863875 | + | 2863767 | 2863783 | -34 | -19 | + | 0.794934 | ACGAACACCGTGTGATT |
| chr5 | 180597576 | 180597725 | + | 180597592 | 180597608 | -59 | -44 | - | 0.794353 | CAGCACACACTGGCCCC |
| chr5 | 180597526 | 180597675 | + | 180597592 | 180597608 | -9 | 6 | - | 0.794353 | CAGCACACACTGGCCCC |
| chr5 | 180597476 | 180597625 | + | 180597592 | 180597608 | 41 | 56 | - | 0.794353 | CAGCACACACTGGCCCC |
| chr19 | 5455226 | 5455375 | + | 5455365 | 5455381 | 64 | 79 | + | 0.792216 | TGGGACAGAATGTAGCA |
| chr12 | 322876 | 323025 | + | 322897 | 322913 | -54 | -39 | + | 0.790961 | GGGCACCCCTTGTCCTG |
| chr12 | 322826 | 322975 | + | 322897 | 322913 | -4 | 11 | + | 0.790961 | GGGCACCCCTTGTCCTG |
| chr6 | 168629976 | 168630125 | + | 168629993 | 168630009 | -58 | -43 | + | 0.790707 | TGGAACCAAGAGTTCCC |
| chr6 | 168629926 | 168630075 | + | 168629993 | 168630009 | -8 | 7 | + | 0.790707 | TGGAACCAAGAGTTCCC |
| chr6 | 168629876 | 168630025 | + | 168629993 | 168630009 | 42 | 57 | + | 0.790707 | TGGAACCAAGAGTTCCC |
| chr2 | 26788626 | 26788775 | + | 26788750 | 26788766 | 49 | 64 | + | 0.789367 | GGTTCCATGCTGTTCCT |
| chr17 | 26795176 | 26795325 | + | 26795229 | 26795245 | -22 | -7 | + | 0.785916 | GAGGACTTGGTCTTCCC |
| chr7 | 89747876 | 89748025 | + | 89747885 | 89747901 | -66 | -51 | - | 0.78478 | CAGTAAACGCTGTCCCG |
| chr7 | 89747826 | 89747975 | + | 89747885 | 89747901 | -16 | -1 | - | 0.78478 | CAGTAAACGCTGTCCCG |
| chr22 | 43827726 | 43827875 | + | 43827739 | 43827755 | -62 | -47 | - | 0.784424 | CAGCACATAATGAGCTC |
| chr3 | 194097026 | 194097175 | + | 194097117 | 194097133 | 16 | 31 | - | 0.783168 | TGGAACACTGGGACCCT |
| chr3 | 194096976 | 194097125 | + | 194097117 | 194097133 | 66 | 81 | - | 0.783168 | TGGAACACTGGGACCCT |
| chr19 | 35981426 | 35981575 | + | 35981490 | 35981506 | -11 | 4 | + | 0.7821 | AGGCATGGACTGTTCCT |
| chr19 | 35981376 | 35981525 | + | 35981490 | 35981506 | 39 | 54 | + | 0.7821 | AGGCATGGACTGTTCCT |
| chr10 | 3568726 | 3568875 | + | 3568746 | 3568762 | -55 | -40 | - | 0.782035 | AAGAACTAAGTGTTACT |
| chr19 | 36195326 | 36195475 | + | 36195411 | 36195427 | 10 | 25 | - | 0.78194 | GGGAGCGTTGCGTCCCT |
| chr19 | 36195276 | 36195425 | + | 36195411 | 36195427 | 60 | 75 | - | 0.78194 | GGGAGCGTTGCGTCCCT |
| chr10 | 131357076 | 131357225 | + | 131357194 | 131357210 | 43 | 58 | + | 0.781329 | GGGGGCGGCGTGTGCTC |
| chr9 | 33448176 | 33448325 | + | 33448220 | 33448236 | -31 | -16 | + | 0.781194 | GAGCACGTCGTGTGTCG |
| chr7 | 1251226 | 1251375 | + | 1251295 | 1251311 | -6 | 9 | - | 0.780791 | TGGTCCACAGGGTCCCT |
| chr7 | 1251176 | 1251325 | + | 1251295 | 1251311 | 44 | 59 | - | 0.780791 | TGGTCCACAGGGTCCCT |
| chr4 | 1564076 | 1564225 | + | 1564145 | 1564161 | -6 | 9 | + | 0.780756 | GAGGACAGAGTGGCCCA |
| chr4 | 1564026 | 1564175 | + | 1564145 | 1564161 | 44 | 59 | + | 0.780756 | GAGGACAGAGTGGCCCA |
| chr22 | 37499726 | 37499875 | + | 37499799 | 37499815 | -2 | 13 | + | 0.780437 | AAGTACATGGGGTGCAG |
| chr10 | 131744276 | 131744425 | + | 131744286 | 131744302 | -65 | -50 | - | 0.780002 | AAAAAAATCCTGTTCCT |
| chr19 | 17138776 | 17138925 | + | 17138916 | 17138932 | 65 | 80 | - | 0.779974 | TCCAACACCCTGTTCTT |
| chr4 | 3752176 | 3752325 | + | 3752188 | 3752204 | -63 | -48 | - | 0.778441 | GGGCTCTCCTTGTTCTT |
| chr22 | 43621726 | 43621875 | + | 43621831 | 43621847 | 30 | 45 | - | 0.777512 | ATGAACAACGCTTTCCT |
| chr17 | 21278776 | 21278925 | + | 21278779 | 21278795 | -72 | -57 | - | 0.777355 | GGGCACAGACTGGCCTC |
| chr10 | 94448376 | 94448525 | + | 94448391 | 94448407 | -60 | -45 | + | 0.776906 | TAGACCATTATGTACAT |
| chr16 | 90114976 | 90115125 | + | 90114994 | 90115010 | -57 | -42 | - | 0.776834 | AATAACACATTCTTCCT |
| chr2 | 54560476 | 54560625 | + | 54560535 | 54560551 | -16 | -1 | + | 0.776389 | GCGAACAGCATGTGCAG |
| chr4 | 74847726 | 74847875 | + | 74847760 | 74847776 | -41 | -26 | - | 0.775917 | CGGCACGTTTTATTCCC |
| chr17 | 60214476 | 60214625 | + | 60214519 | 60214535 | -32 | -17 | - | 0.775747 | CGGCACACCTGGTTCAT |
| chr21 | 47399476 | 47399625 | + | 47399559 | 47399575 | 8 | 23 | - | 0.775201 | TTACACAACGTGTTCCT |
| chr1 | 2424626 | 2424775 | + | 2424666 | 2424682 | -35 | -20 | - | 0.773081 | TGGGTCAGCGCGTGCCT |
| chr1 | 2424576 | 2424725 | + | 2424666 | 2424682 | 15 | 30 | - | 0.773081 | TGGGTCAGCGCGTGCCT |
| chr1 | 2424526 | 2424675 | + | 2424666 | 2424682 | 65 | 80 | - | 0.773081 | TGGGTCAGCGCGTGCCT |
| chr5 | 137225226 | 137225375 | + | 137225274 | 137225290 | -27 | -12 | - | 0.771242 | AGGAACTCTGCGGCCCC |
| chr5 | 137225176 | 137225325 | + | 137225274 | 137225290 | 23 | 38 | - | 0.771242 | AGGAACTCTGCGGCCCC |
| chr17 | 79109726 | 79109875 | + | 79109790 | 79109806 | -11 | 4 | - | 0.77075 | AAGGACACCCTGCACAC |
| chr17 | 79109676 | 79109825 | + | 79109790 | 79109806 | 39 | 54 | - | 0.77075 | AAGGACACCCTGCACAC |
| chr13 | 31620276 | 31620425 | + | 31620408 | 31620424 | 57 | 72 | - | 0.770562 | AGGCCCATCGGGTGCCT |
| chr14 | 106208326 | 106208475 | + | 106208348 | 106208364 | -53 | -38 | + | 0.770025 | CGGTACGTGCTGTTGTA |
| chr14 | 106208276 | 106208425 | + | 106208348 | 106208364 | -3 | 12 | + | 0.770025 | CGGTACGTGCTGTTGTA |
| chr11 | 72980776 | 72980925 | + | 72980865 | 72980881 | 14 | 29 | - | 0.769534 | CGGAACAAGCTTTCCCA |
| chr11 | 72980726 | 72980875 | + | 72980865 | 72980881 | 64 | 79 | - | 0.769534 | CGGAACAAGCTTTCCCA |
| chr7 | 564176 | 564325 | + | 564182 | 564198 | -69 | -54 | + | 0.76916 | GGGGACACGGCGTTAAG |
| chr16 | 1159776 | 1159925 | + | 1159853 | 1159869 | 2 | 17 | + | 0.768869 | GATCACACGCTGATCCT |
| chr16 | 1159726 | 1159875 | + | 1159853 | 1159869 | 52 | 67 | + | 0.768869 | GATCACACGCTGATCCT |

Supplementary Table. 5.2.3. (2/10)
Functional enrichment of principal component three top 1000 segments

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| chr1 | 223435676 | 223435825 | + | 223435710 | 223435726 | -41 | -26 | - | 0.768226 GTGCAAATCGTGTGCTT |
| chr1 | 223435626 | 223435775 | + | 223435710 | 223435726 | 9 | 24 | - | 0.768226 GTGCAAATCGTGTGCTT |
| chr8 | 54164576 | 54164725 | + | 54164615 | 54164631 | -36 | -21 | + | 0.766828 AGTAACAAAATGATCCA |
| chr8 | 54164526 | 54164675 | + | 54164615 | 54164631 | 14 | 29 | + | 0.766828 AGTAACAAAATGATCCA |
| chr16 | 57916776 | 57916925 | + | 57916899 | 57916915 | 48 | 63 | + | 0.76667 AGGCAGAGGATGCTCCT |
| chr19 | 56914726 | 56914875 | + | 56914807 | 56914823 | 6 | 21 | - | 0.766002 CAGAACTCACTGAACTC |
| chr19 | 56914676 | 56914825 | + | 56914807 | 56914823 | 56 | 71 | - | 0.766002 CAGAACTCACTGAACTC |
| chr22 | 25160776 | 25160925 | + | 25160855 | 25160871 | 4 | 19 | + | 0.765194 AGGAACAGTGTGGTTTG |
| chr14 | 89881626 | 89881775 | + | 89881628 | 89881644 | -73 | -58 | + | 0.765096 GACAACATGCTGCTCTC |
| chr7 | 158800876 | 158801025 | + | 158800989 | 158801005 | 38 | 53 | - | 0.764857 TGGTACTTTCTGTATTT |
| chr21 | 44494676 | 44494825 | + | 44494747 | 44494763 | -4 | 11 | - | 0.763198 CAGAAAGCCGTGGACCC |
| chr21 | 44494626 | 44494775 | + | 44494747 | 44494763 | 46 | 61 | - | 0.763198 CAGAAAGCCGTGGACCC |
| chr8 | 142452376 | 142452525 | + | 142452500 | 142452516 | 49 | 64 | - | 0.763149 GGTGTCACCGTGTGCTT |
| chr12 | 111136976 | 111137125 | + | 111137035 | 111137051 | -16 | -1 | + | 0.762786 GGGCACATTTCGTGTTC |
| chr3 | 64225026 | 64225175 | + | 64225123 | 64225139 | 22 | 37 | + | 0.761576 GGGAACATGGAGATGCT |
| chr3 | 196515476 | 196515625 | + | 196515525 | 196515541 | -26 | -11 | - | 0.761283 GTGCACAACATGTGTCC |
| chr3 | 128724926 | 128725075 | + | 128724958 | 128724974 | -43 | -28 | + | 0.761114 AGTTAAGTAATGTACCC |
| chr3 | 128724876 | 128725025 | + | 128724958 | 128724974 | 7 | 22 | + | 0.761114 AGTTAAGTAATGTACCC |
| chr9 | 122734526 | 122734675 | + | 122734566 | 122734582 | -35 | -20 | - | 0.761051 CAGAAGAAATTGTGCTT |
| chr9 | 122734476 | 122734625 | + | 122734566 | 122734582 | 15 | 30 | - | 0.761051 CAGAAGAAATTGTGCTT |
| chr2 | 242797776 | 242797925 | + | 242797796 | 242797812 | -55 | -40 | - | 0.760926 CAGGACATTGCTTTCCA |
| chr2 | 242797726 | 242797875 | + | 242797796 | 242797812 | -5 | 10 | - | 0.760926 CAGGACATTGCTTTCCA |
| chr2 | 242797676 | 242797825 | + | 242797796 | 242797812 | 45 | 60 | - | 0.760926 CAGGACATTGCTTTCCA |
| chr2 | 27938076 | 27938225 | + | 27938200 | 27938216 | 49 | 64 | + | 0.760802 GGGGACACCATGGTGCA |
| chr5 | 137225326 | 137225475 | + | 137225357 | 137225373 | -44 | -29 | + | 0.76041 AGGGACAGATGGTCCCA |
| chr5 | 137225276 | 137225425 | + | 137225357 | 137225373 | 6 | 21 | + | 0.76041 AGGGACAGATGGTCCCA |
| chr7 | 1329376 | 1329525 | + | 1329404 | 1329420 | -47 | -32 | + | 0.760336 GGGGACGGGGGGGAACCC |
| chr7 | 1329326 | 1329475 | + | 1329404 | 1329420 | 3 | 18 | + | 0.760336 GGGGACGGGGGGGAACCC |
| chr7 | 1329276 | 1329425 | + | 1329404 | 1329420 | 53 | 68 | + | 0.760336 GGGGACGGGGGGGAACCC |
| chr1 | 3347876 | 3348025 | + | 3347983 | 3347999 | 32 | 47 | + | 0.760279 TGGGACAGCGCGTGCAG |
| chr3 | 196515526 | 196515675 | + | 196515643 | 196515659 | 42 | 57 | + | 0.760258 GGGAAGAGCCTGTTTCA |
| chr14 | 93154676 | 93154825 | + | 93154804 | 93154820 | 53 | 68 | - | 0.760198 CGAGACGGCATGTTCTT |
| chr3 | 194090526 | 194090675 | + | 194090660 | 194090676 | 59 | 74 | - | 0.75894 CAGGACACATTTTCCTC |
| chr1 | 204655126 | 204655275 | + | 204655190 | 204655206 | -11 | 4 | + | 0.75865 AGTCACAGCATGCACCT |
| chr1 | 204655076 | 204655225 | + | 204655190 | 204655206 | 39 | 54 | + | 0.75865 AGTCACAGCATGCACCT |
| chr22 | 50457076 | 50457225 | + | 50457172 | 50457188 | 21 | 36 | - | 0.758632 GGCTGCAACCTGTACTC |
| chr22 | 37771226 | 37771375 | + | 37771275 | 37771291 | -26 | -11 | - | 0.758478 AGGAAGGCACCGTCCTC |
| chr22 | 37771176 | 37771325 | + | 37771275 | 37771291 | 24 | 39 | - | 0.758478 AGGAAGGCACCGTCCTC |
| chr14 | 105105026 | 105105175 | + | 105105144 | 105105160 | 43 | 58 | + | 0.758448 GAGAACACAGCGTAGGG |
| chr2 | 159705526 | 159705675 | + | 159705566 | 159705582 | -35 | -20 | - | 0.756755 AAGTAACCCCTGTGCTT |
| chr2 | 159705476 | 159705625 | + | 159705566 | 159705582 | 15 | 30 | - | 0.756755 AAGTAACCCCTGTGCTT |
| chr5 | 111090026 | 111090175 | + | 111090131 | 111090147 | 30 | 45 | + | 0.756378 AAGTAGGTAACGTTCTG |
| chr15 | 99974726 | 99974875 | + | 99974826 | 99974842 | 25 | 40 | - | 0.756 CAGAACACGGTGACGCT |
| chr14 | 94406426 | 94406575 | + | 94406556 | 94406572 | 55 | 70 | + | 0.75556 AGGGACAGCCAGATCCC |
| chr17 | 74237526 | 74237675 | + | 74237567 | 74237583 | -34 | -19 | - | 0.755025 TAGAACATGACTTTCAC |
| chr16 | 14380576 | 14380725 | + | 14380621 | 14380637 | -30 | -15 | - | 0.754879 CGCCACATCGTGCACTC |
| chr16 | 14380526 | 14380675 | + | 14380621 | 14380637 | 20 | 35 | - | 0.754879 CGCCACATCGTGCACTC |
| chr12 | 47219876 | 47220025 | + | 47219960 | 47219976 | 9 | 24 | + | 0.754653 GTGTACCCAGTGTTATT |
| chr10 | 3373226 | 3373375 | + | 3373332 | 3373348 | 31 | 46 | + | 0.754633 GAGAGAAGGCTGTTCTC |
| chr8 | 914376 | 914525 | + | 914488 | 914504 | 37 | 52 | - | 0.753319 TCTAACATGCTGTGCCT |
| chr4 | 3288776 | 3288925 | + | 3288790 | 3288806 | -61 | -46 | + | 0.753066 AGGGGCACCTAGTTCTT |
| chr4 | 3288726 | 3288875 | + | 3288790 | 3288806 | -11 | 4 | + | 0.753066 AGGGGCACCTAGTTCTT |
| chr4 | 3288676 | 3288825 | + | 3288790 | 3288806 | 39 | 54 | + | 0.753066 AGGGGCACCTAGTTCTT |
| chr3 | 185788626 | 185788775 | + | 185788626 | 185788642 | -75 | -60 | - | 0.752785 AGGCACAGCCTGGTTCC |
| chr16 | 29242026 | 29242175 | + | 29242075 | 29242091 | -26 | -11 | - | 0.752636 GGCTATAGCCTGTGCCC |
| chr16 | 29241976 | 29242125 | + | 29242075 | 29242091 | 24 | 39 | - | 0.752636 GGCTATAGCCTGTGCCC |
| chr7 | 120967726 | 120967875 | + | 120967734 | 120967750 | -67 | -52 | + | 0.752458 GGGTACCACTGAGCTG |
| chr19 | 41061926 | 41062075 | + | 41061926 | 41061942 | -75 | -60 | - | 0.751816 GGGGTCACGCTCTGCCC |
| chr9 | 137731726 | 137731875 | + | 137731836 | 137731852 | 35 | 50 | + | 0.751691 TGGAACTGTGTGTTTTG |
| chr1 | 210612226 | 210612375 | + | 210612353 | 210612369 | 52 | 67 | - | 0.75115 AGGAAGGCCATGTGGCT |
| chr14 | 105044926 | 105045075 | + | 105044929 | 105044945 | -72 | -57 | + | 0.75114 GAGAAAAGTTTGCTCCC |
| chr1 | 205913876 | 205914025 | + | 205913951 | 205913967 | 0 | 15 | + | 0.750646 GGGAAGGATGTGTGCAT |
| chr1 | 205913826 | 205913975 | + | 205913951 | 205913967 | 50 | 65 | + | 0.750646 GGGAAGGATGTGTGCAT |
| chr17 | 25798576 | 25798725 | + | 25798580 | 25798596 | -71 | -56 | - | 0.750629 TGGAACACAGTGTCAGG |
| chr22 | 37771276 | 37771425 | + | 37771296 | 37771312 | -55 | -40 | - | 0.749976 TAGGAGATCTCGTTCTT |
| chr6 | 169351276 | 169351425 | + | 169351296 | 169351312 | -55 | -40 | - | 0.749828 GAGAGCAACTTTTACCT |
| chr4 | 3776376 | 3776525 | + | 3776466 | 3776482 | 15 | 30 | + | 0.747761 GATTACACGTTGAGCCA |
| chr1 | 156828576 | 156828725 | + | 156828711 | 156828727 | 60 | 75 | + | 0.746071 AATGACACACAGGGTTCTG |
| chr10 | 131691176 | 131691325 | + | 131691239 | 131691255 | -12 | 3 | - | 0.743974 TGGCACACACGGTTCGG |
| chr5 | 172924726 | 172924875 | + | 172924729 | 172924745 | -72 | -57 | - | 0.74366 TAGGACAATGCCTTCCA |
| chr9 | 135033126 | 135033275 | + | 135033195 | 135033211 | -6 | 9 | - | 0.743455 AGGCACGACTCATACCC |
| chr10 | 4697326 | 4697475 | + | 4697372 | 4697388 | -29 | -14 | - | 0.743364 CAGGACAGAACATTCTC |
| chr10 | 4697276 | 4697425 | + | 4697372 | 4697388 | 21 | 36 | - | 0.743364 CAGGACAGAACATTCTC |
| chr10 | 3494226 | 3494375 | + | 3494241 | 3494257 | -60 | -45 | - | 0.742628 CATTCCACATTGTTCTT |
| chr2 | 242797826 | 242797975 | + | 242797848 | 242797864 | -53 | -38 | + | 0.742396 AGGAACCTGCTGGCCCA |
| chr6 | 33561326 | 33561475 | + | 33561447 | 33561463 | 46 | 61 | + | 0.742236 GCGAACAGTCTGTGATT |
| chr3 | 97542076 | 97542225 | + | 97542087 | 97542103 | -64 | -49 | - | 0.742191 GCTTACACTGTCTTCTC |
| chr10 | 12543276 | 12543425 | + | 12543287 | 12543303 | -64 | -49 | + | 0.742054 CAGAGCACTGTGCACTG |
| chr14 | 104865726 | 104865875 | + | 104865747 | 104865763 | -54 | -39 | - | 0.742026 GAGTAAGACATCTTCCC |
| chr1 | 3534326 | 3534475 | + | 3534415 | 3534431 | 14 | 29 | + | 0.741944 GGGAACCGGGGGTTGCT |
| chr1 | 3534276 | 3534425 | + | 3534415 | 3534431 | 64 | 79 | + | 0.741944 GGGAACCGGGGGTTGCT |
| chr17 | 21278826 | 21278975 | + | 21278844 | 21278860 | -57 | -42 | + | 0.74184 GACAAAATGGAGTTCCC |
| chr20 | 30618776 | 30618925 | + | 30618826 | 30618842 | -25 | -10 | - | 0.741588 CAGGACATCGGCTACTT |
| chr7 | 99987426 | 99987575 | + | 99987456 | 99987472 | -45 | -30 | - | 0.740949 CCAGACACCCTGTGCCC |
| chr9 | 140127226 | 140127375 | + | 140127258 | 140127274 | -43 | -28 | - | 0.740884 AAGGACAACGTGGTGCT |
| chr9 | 140127176 | 140127325 | + | 140127258 | 140127274 | 7 | 22 | - | 0.740884 AAGGACAACGTGGTGCT |
| chr12 | 133178876 | 133179025 | + | 133178940 | 133178956 | -11 | 4 | - | 0.740797 CCTAACACTGTGTTCAT |
| chr10 | 3448926 | 3449075 | + | 3449026 | 3449042 | 25 | 40 | + | 0.740721 AGGCACCAGCGGTTCCA |
| chr12 | 299626 | 299775 | + | 299638 | 299654 | -63 | -48 | + | 0.740254 AAGAACAGTGCGGAACT |
| chr1 | 23521276 | 23521425 | + | 23521301 | 23521317 | -50 | -35 | - | 0.740157 CAGCTCGCGGTTTTCCC |
| chr8 | 1644876 | 1645025 | + | 1644898 | 1644914 | -53 | -38 | - | 0.739654 AGGTATAGAATGGGCTC |
| chr8 | 1644826 | 1644975 | + | 1644898 | 1644914 | -3 | 12 | - | 0.739654 AGGTATAGAATGGGCTC |
| chr7 | 1266026 | 1266175 | + | 1266161 | 1266177 | 60 | 75 | - | 0.738836 CTGAACAACGTGAGCCG |
| chr15 | 29825226 | 29825375 | + | 29825252 | 29825268 | -49 | -34 | - | 0.737961 TTGTACACAGGCTACTT |
| chr17 | 74581326 | 74581475 | + | 74581450 | 74581466 | 49 | 64 | - | 0.737698 CAATAGAAAGTGTACTT |
| chr2 | 3697376 | 3697525 | + | 3697391 | 3697407 | -60 | -45 | + | 0.73762 AGGAAACCAGAGTGCCC |
| chr16 | 89009426 | 89009575 | + | 89009486 | 89009502 | -15 | 0 | + | 0.737386 AGGAAAACATTCTGCCG |
| chr16 | 89009376 | 89009525 | + | 89009486 | 89009502 | 35 | 50 | + | 0.737386 AGGAAAACATTCTGCCG |
| chr17 | 76522776 | 76522925 | + | 76522837 | 76522853 | -14 | 1 | + | 0.737288 AAGTACAGAGAGTCACG |
| chr1 | 200143026 | 200143175 | + | 200143117 | 200143133 | 16 | 31 | + | 0.737135 TGGTAGAAGGTGTCCAG |
| chr7 | 1251126 | 1251275 | + | 1251231 | 1251247 | 30 | 45 | - | 0.737062 AAGAGAACTCCGTTCTC |

Supplementary Table. 5.2.3. (3/10)
Functional enrichment of principal component three top 1000 segments

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| chr20 | 43377476 | 43377625 | + | 43377484 | 43377500 | -67 | -52 | - | 0.736943 TAGCACACAGCCATCCC |
| chr20 | 43377426 | 43377575 | + | 43377484 | 43377500 | -17 | -2 | - | 0.736943 TAGCACACAGCCATCCC |
| chr3 | 14862826 | 14862975 | + | 14862861 | 14862877 | -40 | -25 | - | 0.736523 CGGTCCATCTCCTTCCC |
| chr1 | 94270076 | 94270225 | + | 94270197 | 94270213 | 46 | 61 | + | 0.736275 CAGAACAGGATGCAATT |
| chr8 | 143535776 | 143535925 | + | 143535781 | 143535797 | -70 | -55 | + | 0.735991 GAGCACACCGTGAGGTG |
| chr8 | 143535726 | 143535875 | + | 143535781 | 143535797 | -20 | -5 | + | 0.735991 GAGCACACCGTGAGGTG |
| chr8 | 143535676 | 143535825 | + | 143535781 | 143535797 | 30 | 45 | + | 0.735991 GAGCACACCGTGAGGTG |
| chr8 | 143535626 | 143535775 | + | 143535711 | 143535727 | 10 | 25 | - | 0.735584 CGGAACACCAGGGTGCC |
| chr8 | 143535576 | 143535725 | + | 143535711 | 143535727 | 60 | 75 | - | 0.735584 CGGAACACCAGGGTGCC |
| chr1 | 226756376 | 226756525 | + | 226756515 | 226756531 | 64 | 79 | + | 0.735218 AGGGACAGACCTTGCTC |
| chr11 | 67462776 | 67462925 | + | 67462837 | 67462853 | -14 | 1 | + | 0.734221 GTGAACAGAGTTTATCT |
| chr11 | 67462726 | 67462875 | + | 67462837 | 67462853 | 36 | 51 | + | 0.734221 GTGAACAGAGTTTATCT |
| chr14 | 106174226 | 106174375 | + | 106174228 | 106174244 | -73 | -58 | - | 0.73318 CTGGACACGCTGTAGCA |
| chr14 | 106174176 | 106174325 | + | 106174228 | 106174244 | -23 | -8 | - | 0.73318 CTGGACACGCTGTAGCA |
| chr14 | 106174126 | 106174275 | + | 106174228 | 106174244 | 27 | 42 | - | 0.73318 CTGGACACGCTGTAGCA |
| chr8 | 144471976 | 144472125 | + | 144472092 | 144472108 | 41 | 56 | + | 0.732895 GAGGGCCTGGTGTGCTG |
| chr9 | 89410926 | 89411075 | + | 89410955 | 89410971 | -46 | -31 | + | 0.732773 AGGAGCGTGTTCTTCCA |
| chr9 | 89410876 | 89411025 | + | 89410955 | 89410971 | 4 | 19 | + | 0.732773 AGGAGCGTGTTCTTCCA |
| chr9 | 89410826 | 89410975 | + | 89410955 | 89410971 | 54 | 69 | + | 0.732773 AGGAGCGTGTTCTTCCA |
| chr13 | 106063076 | 106063225 | + | 106063080 | 106063096 | -71 | -56 | - | 0.732705 GAGTACATGCCTTTTTC |
| chr13 | 106063026 | 106063175 | + | 106063080 | 106063096 | -21 | -6 | - | 0.732705 GAGTACATGCCTTTTTC |
| chr13 | 106062976 | 106063125 | + | 106063080 | 106063096 | 29 | 44 | - | 0.732705 GAGTACATGCCTTTTTC |
| chr14 | 100624926 | 100625075 | + | 100624984 | 100625000 | -17 | -2 | + | 0.73183 GATTACAAAGTGTGGCA |
| chr14 | 100624876 | 100625025 | + | 100624984 | 100625000 | 33 | 48 | + | 0.73183 GATTACAAAGTGTGGCA |
| chr10 | 130844126 | 130844275 | + | 130844182 | 130844198 | -19 | -4 | + | 0.731576 GGCGACATTTAGTGCCC |
| chr4 | 3691226 | 3691375 | + | 3691265 | 3691281 | -36 | -21 | - | 0.7315 TGGTGCACTGCTGCTT |
| chr11 | 2292976 | 2293125 | + | 2293042 | 2293058 | -9 | 6 | + | 0.731325 TGGGACGTCGAGAACTT |
| chr19 | 14313576 | 14313725 | + | 14313654 | 14313670 | 3 | 18 | - | 0.731231 CAGCACAGGAGGGACCT |
| chr10 | 3343026 | 3343175 | + | 3343038 | 3343054 | -63 | -48 | + | 0.731039 AGAAACAGCGAGTAGCC |
| chr10 | 3342976 | 3343125 | + | 3343038 | 3343054 | -13 | 2 | + | 0.731039 AGAAACAGCGAGTAGCC |
| chr10 | 3342926 | 3343075 | + | 3343038 | 3343054 | 37 | 52 | + | 0.731039 AGAAACAGCGAGTAGCC |
| chr22 | 25160376 | 25160525 | + | 25160443 | 25160459 | -8 | 7 | - | 0.731033 TGGTAAGCACTGATCTC |
| chr11 | 75222326 | 75222475 | + | 75222344 | 75222360 | -57 | -42 | + | 0.731003 AAGGTCAGGGTTTGCCC |
| chr4 | 1535526 | 1535675 | + | 1535526 | 1535542 | -75 | -60 | - | 0.730619 GAGGCCACCATCTGCTT |
| chr1 | 226756326 | 226756475 | + | 226756329 | 226756345 | -72 | -57 | + | 0.730596 AAGAACAGCCAGGTCTG |
| chr11 | 92806326 | 92806475 | + | 92806352 | 92806368 | -49 | -34 | - | 0.73058 CAGCACTCCCCATACCC |
| chr14 | 94463726 | 94463875 | + | 94463846 | 94463862 | 45 | 60 | - | 0.730398 AACTAGACCATTTTCCC |
| chr11 | 45392476 | 45392625 | + | 45392606 | 45392622 | 55 | 70 | - | 0.730197 GAATACTTCCTGTCCTA |
| chr19 | 36004876 | 36005025 | + | 36004923 | 36004939 | -28 | -13 | + | 0.729784 GGGGACAGGACCCACCC |
| chr19 | 36004826 | 36004975 | + | 36004923 | 36004939 | 22 | 37 | + | 0.729784 GGGGACAGGACCCACCC |
| chr10 | 131706676 | 131706825 | + | 131706792 | 131706808 | 41 | 56 | + | 0.729328 GGAGACGGGATATACCT |
| chr10 | 4194376 | 4194525 | + | 4194404 | 4194420 | -47 | -32 | - | 0.728399 AGGAACAGCATGGCGCA |
| chr12 | 86230726 | 86230875 | + | 86230746 | 86230762 | -55 | -40 | + | 0.728111 AGGACCACAGGGCTCCA |
| chr12 | 86230676 | 86230825 | + | 86230746 | 86230762 | -5 | 10 | + | 0.728111 AGGACCACAGGGCTCCA |
| chr14 | 103569326 | 103569475 | + | 103569374 | 103569390 | -27 | -12 | - | 0.727961 GGTTCCACCAGGTGCCT |
| chr14 | 103569276 | 103569425 | + | 103569374 | 103569390 | 23 | 38 | - | 0.727961 GGTTCCACCAGGTGCCT |
| chr10 | 3300426 | 3300575 | + | 3300505 | 3300521 | 4 | 19 | + | 0.727881 GAGGACAAGAGGCTCCG |
| chr10 | 3300376 | 3300525 | + | 3300505 | 3300521 | 54 | 69 | + | 0.727881 GAGGACAAGAGGCTCCG |
| chr10 | 4386776 | 4386925 | + | 4386889 | 4386905 | 38 | 53 | + | 0.727255 TGGAACAGGATGACTCA |
| chr5 | 150538326 | 150538475 | + | 150538362 | 150538378 | -39 | -24 | - | 0.726304 AATAAAACAGTATACCC |
| chr5 | 150538276 | 150538425 | + | 150538362 | 150538378 | 11 | 26 | - | 0.726304 AATAAAACAGTATACCC |
| chr5 | 150538226 | 150538375 | + | 150538362 | 150538378 | 61 | 76 | - | 0.726304 AATAAAACAGTATACCC |
| chr1 | 64197376 | 64197525 | + | 64197467 | 64197483 | 16 | 31 | + | 0.726278 AAGAAGGACGCGTGCTG |
| chr1 | 64197326 | 64197475 | + | 64197467 | 64197483 | 66 | 81 | + | 0.726278 AAGAAGGACGCGTGCTG |
| chr9 | 132482326 | 132482475 | + | 132482328 | 132482344 | -73 | -58 | + | 0.725764 TGGCACCAAGGGTTCTG |
| chr9 | 132482276 | 132482425 | + | 132482328 | 132482344 | -23 | -8 | + | 0.725764 TGGCACCAAGGGTTCTG |
| chr13 | 113807776 | 113807925 | + | 113807807 | 113807823 | -44 | -29 | - | 0.725739 AGGCACAGCTGGGTCCA |
| chr3 | 46622476 | 46622625 | + | 46622516 | 46622532 | -35 | -20 | - | 0.724816 GAGAAGAGCAGGTTCAC |
| chr11 | 120592076 | 120592225 | + | 120592146 | 120592162 | -5 | 10 | + | 0.72459 AGAGATAGTTTGTTCCT |
| chr11 | 120592026 | 120592175 | + | 120592146 | 120592162 | 45 | 60 | + | 0.72459 AGAGATAGTTTGTTCCT |
| chr8 | 142452326 | 142452475 | + | 142452332 | 142452348 | -69 | -54 | + | 0.724217 GAGGACATGGAGAAACC |
| chr11 | 120561176 | 120561325 | + | 120561192 | 120561208 | -59 | -44 | + | 0.723138 AAGAACTGGCTGCGCCA |
| chr9 | 132383026 | 132383175 | + | 132383107 | 132383123 | 6 | 21 | - | 0.722221 AAGTACATGCCAGTCTT |
| chr5 | 149683176 | 149683325 | + | 149683206 | 149683222 | -45 | -30 | + | 0.722198 TGGAACCCGATGAGCAT |
| chr9 | 132383276 | 132383425 | + | 132383381 | 132383397 | 30 | 45 | - | 0.722094 GAGCACTTAGCCTGCCT |
| chr20 | 61979376 | 61979525 | + | 61979434 | 61979450 | -17 | -2 | - | 0.721773 CAGACCAAAGCCTTCCC |
| chr20 | 61979326 | 61979475 | + | 61979434 | 61979450 | 33 | 48 | - | 0.721773 CAGACCAAAGCCTTCCC |
| chr2 | 86037076 | 86037225 | + | 86037201 | 86037217 | 50 | 65 | + | 0.721425 GAATACATGAAGAACCA |
| chr7 | 1458976 | 1459125 | + | 1459060 | 1459076 | 9 | 24 | + | 0.720484 CCCTACGGGGTGTGCCC |
| chr9 | 138171626 | 138171775 | + | 138171767 | 138171783 | 66 | 81 | + | 0.720375 TGGGAGAATGTGTTCGG |
| chr3 | 127173576 | 127173725 | + | 127173601 | 127173617 | -50 | -35 | - | 0.720203 CAGGGCTTGGTGTCCCT |
| chr1 | 157140676 | 157140825 | + | 157140703 | 157140719 | -48 | -33 | - | 0.720167 GGGGGCGCCCTGGACTC |
| chr6 | 17988876 | 17989025 | + | 17988904 | 17988920 | -47 | -32 | - | 0.720053 GTGTACGGCAGGTGCTA |
| chr6 | 17988826 | 17988975 | + | 17988904 | 17988920 | 3 | 18 | - | 0.720053 GTGTACGGCAGGTGCTA |
| chr7 | 2728676 | 2728825 | + | 2728685 | 2728701 | -66 | -51 | + | 0.71978 AAGCCCCCGGTTTTCCC |
| chr10 | 3479976 | 3480125 | + | 3480023 | 3480039 | -28 | -13 | + | 0.719774 GAGAGCGGACTGTTACT |
| chr10 | 3479926 | 3480075 | + | 3480023 | 3480039 | 22 | 37 | + | 0.719774 GAGAGCGGACTGTTACT |
| chr7 | 4843826 | 4843975 | + | 4843860 | 4843876 | -41 | -26 | - | 0.719772 CGGAAGAGTGCGTGCAT |
| chr14 | 104639676 | 104639825 | + | 104639702 | 104639718 | -49 | -34 | + | 0.719625 GAGAAGGCGGCTTTCTA |
| chr20 | 62004676 | 62004825 | + | 62004700 | 62004716 | -51 | -36 | - | 0.719422 CGGGACTGACTCTTCCT |
| chr20 | 62004626 | 62004775 | + | 62004700 | 62004716 | -1 | 14 | - | 0.719422 CGGGACTGACTCTTCCT |
| chr7 | 30029826 | 30029975 | + | 30029863 | 30029879 | -38 | -23 | - | 0.719243 AGCCAGACCCTGTCCCA |
| chr7 | 30029776 | 30029925 | + | 30029863 | 30029879 | 12 | 27 | - | 0.719243 AGCCAGACCCTGTCCCA |
| chr22 | 18508226 | 18508375 | + | 18508262 | 18508278 | -39 | -24 | - | 0.718404 GGGCTCTCTCTGGTCCC |
| chr16 | 24697376 | 24697525 | + | 24697386 | 24697402 | -65 | -50 | - | 0.718384 CAGGACCTCTTGGACCT |
| chr16 | 24697326 | 24697475 | + | 24697386 | 24697402 | -15 | 0 | - | 0.718384 CAGGACCTCTTGGACCT |
| chr1 | 21913376 | 21913525 | + | 21913418 | 21913434 | -33 | -18 | + | 0.718288 GAGTACAGGATTTTAAT |
| chr20 | 23969726 | 23969875 | + | 23969764 | 23969780 | -37 | -22 | + | 0.718285 AGTTCCAGGGCGTCCTC |
| chr7 | 23471726 | 23471875 | + | 23471789 | 23471805 | -12 | 3 | + | 0.717889 TAGAACAGAGTCTGACA |
| chr7 | 23471676 | 23471825 | + | 23471789 | 23471805 | 38 | 53 | + | 0.717889 TAGAACAGAGTCTGACA |
| chr10 | 4230476 | 4230625 | + | 4230509 | 4230525 | -42 | -27 | + | 0.71772 AAGCATCCTGTGAACCC |
| chr10 | 4230426 | 4230575 | + | 4230509 | 4230525 | 8 | 23 | + | 0.71772 AAGCATCCTGTGAACCC |
| chr22 | 26877576 | 26877725 | + | 26877667 | 26877683 | 16 | 31 | - | 0.716807 TGGCAGGCCTTGTTACT |
| chr22 | 26877526 | 26877675 | + | 26877667 | 26877683 | 66 | 81 | - | 0.716807 TGGCAGGCCTTGTTACT |
| chr20 | 44934626 | 44934775 | + | 44934681 | 44934697 | -20 | -5 | - | 0.716715 GGCCAAACTTTGTCCTC |
| chr20 | 44934576 | 44934725 | + | 44934681 | 44934697 | 30 | 45 | - | 0.716715 GGCCAAACTTTGTCCTC |
| chr1 | 2527376 | 2527525 | + | 2527389 | 2527405 | -62 | -47 | - | 0.716309 CGGTGCTGCCCGTGCCC |
| chr1 | 2527326 | 2527475 | + | 2527389 | 2527405 | -12 | 3 | - | 0.716309 CGGTGCTGCCCGTGCCC |
| chr3 | 64305676 | 64305825 | + | 64305708 | 64305724 | -43 | -28 | - | 0.715649 TAGAACTTTCCATACTT |
| chr3 | 64305626 | 64305775 | + | 64305708 | 64305724 | 7 | 22 | - | 0.715649 TAGAACTTTCCATACTT |
| chr14 | 93154626 | 93154775 | + | 93154699 | 93154715 | -2 | 13 | - | 0.715579 CATTCCATGACGTGCCC |

Supplementary Table. 5.2.3. (4/10)
Functional enrichment of principal component three top 1000 segments

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| chr5 | 1010876 | 1011025 | + | 1011007 | 1011023 | 56 | 71 | - | 0.715578 CGCTGCACCATGAACCT |
| chr15 | 62358676 | 62358825 | + | 62358677 | 62358693 | -74 | -59 | + | 0.715567 AGACATGAAATGTTCCC |
| chr6 | 170585776 | 170585925 | + | 170585893 | 170585909 | 42 | 57 | - | 0.715446 TGCAACATACTGACCTT |
| chr19 | 17138726 | 17138875 | + | 17138740 | 17138756 | -61 | -46 | - | 0.715442 AGTCGCACTGTGTTGCC |
| chr9 | 129282626 | 129282775 | + | 129282661 | 129282677 | -40 | -25 | - | 0.715401 AACAAGACCAGGTACTT |
| chr9 | 129282576 | 129282725 | + | 129282661 | 129282677 | 10 | 25 | - | 0.715401 AACAAGACCAGGTACTT |
| chr9 | 139587826 | 139587975 | + | 139587868 | 139587884 | -33 | -18 | - | 0.715137 CTGGACACCCTCTGCCG |
| chr9 | 139587776 | 139587925 | + | 139587868 | 139587884 | 17 | 32 | - | 0.715137 CTGGACACCCTCTGCCG |
| chr14 | 70476726 | 70476875 | + | 70476858 | 70476874 | 57 | 72 | + | 0.71504 AGCAGCATTGTGTAATT |
| chr1 | 200175476 | 200175625 | + | 200175496 | 200175512 | -55 | -40 | - | 0.714357 GACAACACCTTATCCTA |
| chr17 | 27347076 | 27347225 | + | 27347145 | 27347161 | -6 | 9 | + | 0.71427 GGGCCCGAGGTGAGCTT |
| chr17 | 77536126 | 77536275 | + | 77536263 | 77536279 | 62 | 77 | + | 0.714121 AAGAACATTCTGACTCG |
| chr15 | 33437276 | 33437425 | + | 33437376 | 33437392 | 25 | 40 | - | 0.713709 GGCTGCACAAGGTGCCT |
| chr14 | 101123276 | 101123425 | + | 101123308 | 101123324 | -43 | -28 | - | 0.713096 CAGCACCCGCTGTCACT |
| chr14 | 101123226 | 101123375 | + | 101123308 | 101123324 | 7 | 22 | - | 0.713096 CAGCACCCGCTGTCACT |
| chr11 | 118042626 | 118042775 | + | 118042666 | 118042682 | -35 | -20 | + | 0.713089 AGGGACAAGCTGAGGGCC |
| chr5 | 2204476 | 2204625 | + | 2204561 | 2204577 | 10 | 25 | + | 0.713081 GGGCACAGTGGGAATCC |
| chr10 | 4358426 | 4358575 | + | 4358449 | 4358465 | -52 | -37 | - | 0.712507 AAACACATTATGTCTTT |
| chr10 | 3378776 | 3378925 | + | 3378813 | 3378829 | -38 | -23 | - | 0.712474 TTGCACAGCACGCACTT |
| chr16 | 67335976 | 67336125 | + | 67336007 | 67336023 | -44 | -29 | - | 0.712111 CATTACTTTGTTTTCCT |
| chr14 | 104851976 | 104852125 | + | 104852067 | 104852083 | 16 | 31 | - | 0.71207 TAGAACACCAGTCACTC |
| chr14 | 104851926 | 104852075 | + | 104852067 | 104852083 | 66 | 81 | - | 0.71207 TAGAACACCAGTCACTC |
| chr17 | 35277326 | 35277475 | + | 35277377 | 35277393 | -24 | -9 | + | 0.711936 GGGAGCAGGCAGGACCC |
| chr17 | 35277276 | 35277425 | + | 35277377 | 35277393 | 26 | 41 | + | 0.711936 GGGAGCAGGCAGGACCC |
| chr1 | 231761526 | 231761675 | - | 231761653 | 231761669 | 52 | 67 | - | 0.71168 AGGACCTCAGCCTTCCT |
| chr11 | 62100676 | 62100825 | + | 62100695 | 62100711 | -56 | -41 | + | 0.710832 GGAAAGGTGGTTTTCCC |
| chr11 | 62100626 | 62100775 | + | 62100695 | 62100711 | -6 | 9 | + | 0.710832 GGAAAGGTGGTTTTCCC |
| chr11 | 62100576 | 62100725 | + | 62100695 | 62100711 | 44 | 59 | + | 0.710832 GGAAAGGTGGTTTTCCC |
| chr16 | 49530476 | 49530625 | + | 49530499 | 49530515 | -52 | -37 | - | 0.710661 AATAAAACTGTTTTCTT |
| chr3 | 54353576 | 54353725 | + | 54353643 | 54353659 | -8 | 7 | + | 0.710601 AGTTAAACCGTGTTTTG |
| chr19 | 554876 | 555025 | + | 554944 | 554960 | -7 | 8 | - | 0.710297 TGGCACGTCCTGGTGCT |
| chr19 | 554826 | 554975 | + | 554944 | 554960 | 43 | 58 | - | 0.710297 TGGCACGTCCTGGTGCT |
| chr14 | 106095426 | 106095575 | + | 106095533 | 106095549 | 32 | 47 | - | 0.710252 CATCCGTCATGTTCCT |
| chr1 | 7130326 | 7130475 | + | 7130462 | 7130478 | 61 | 76 | - | 0.710251 AGTTACTTGTTTTTCCT |
| chr10 | 5406476 | 5406625 | + | 5406607 | 5406623 | 56 | 71 | + | 0.710223 AGGTCCAGGCTCTACAT |
| chr4 | 7652026 | 7652175 | + | 7652086 | 7652102 | -15 | 0 | - | 0.709356 CAGAACGTGCTGTGTGC |
| chr8 | 144303226 | 144303375 | + | 144303297 | 144303313 | -4 | 11 | - | 0.708927 AGTTTCATTTTGCTCCC |
| chr8 | 144303176 | 144303325 | + | 144303297 | 144303313 | 46 | 61 | - | 0.708927 AGTTTCATTTTGCTCCC |
| chr4 | 1537476 | 1537625 | + | 1537489 | 1537505 | -62 | -47 | + | 0.708447 ATGTGCACAATGTGCAG |
| chr1 | 200003226 | 200003375 | + | 200003308 | 200003324 | 7 | 22 | - | 0.708441 AGTTTCACAAGGTGCTT |
| chr4 | 3288826 | 3288975 | + | 3288926 | 3288942 | 25 | 40 | - | 0.708372 TTGTTCAAGGTGGACTC |
| chr17 | 81035976 | 81036125 | + | 81035996 | 81036012 | -55 | -40 | + | 0.708355 TGTCACTTGATGTGCTG |
| chr5 | 10653226 | 10653375 | + | 10653307 | 10653323 | 6 | 21 | + | 0.708336 GGGTCCAGGCAGCTCCT |
| chr5 | 10653176 | 10653325 | + | 10653307 | 10653323 | 56 | 71 | + | 0.708336 GGGTCCAGGCAGCTCCT |
| chr10 | 3250676 | 3250825 | + | 3250793 | 3250809 | 42 | 57 | - | 0.707985 GGGTGCGCTGCTTCCTC |
| chr8 | 25902126 | 25902275 | + | 25902156 | 25902172 | -45 | -30 | + | 0.707899 AGGCACGGGGTGCGCGG |
| chr10 | 94448476 | 94448625 | + | 94448587 | 94448603 | 36 | 51 | + | 0.707897 TTGGACAAATAGTTCCA |
| chr1 | 21044826 | 21044975 | + | 21044850 | 21044866 | -51 | -36 | + | 0.707717 AAGGACCTGTTGCCCCT |
| chr19 | 17571526 | 17571675 | + | 17571661 | 17571677 | 60 | 75 | + | 0.707689 GAGAACAGGGAGGCCAT |
| chr7 | 73465976 | 73466125 | + | 73466112 | 73466128 | 61 | 76 | + | 0.707666 GAGGGGCTGGTGTTCCT |
| chr6 | 168533376 | 168533525 | + | 168533434 | 168533450 | -17 | -2 | + | 0.707438 AAGATCAGCGCGTCCAG |
| chr5 | 114514576 | 114514725 | + | 114514686 | 114514702 | 35 | 50 | + | 0.707295 GGCTACAGAATCATCCT |
| chr1 | 7601876 | 7602025 | + | 7601897 | 7601913 | -54 | -39 | + | 0.706944 GGGCTCAGAATGCGCTG |
| chr1 | 15655776 | 15655925 | + | 15655812 | 15655828 | -39 | -24 | + | 0.706872 GGGCTCCTCATGATCTC |
| chr14 | 104617876 | 104618025 | + | 104617905 | 104617921 | -46 | -31 | + | 0.706808 AGGGAGAGGGTCTCCTG |
| chr14 | 104617826 | 104617975 | + | 104617905 | 104617921 | 4 | 19 | + | 0.706808 AGGGAGAGGGTCTCCTG |
| chr12 | 124607826 | 124607975 | + | 124607931 | 124607947 | 30 | 45 | + | 0.706445 ATGAACCCTTTATTCCA |
| chr3 | 14595776 | 14595925 | + | 14595898 | 14595914 | 47 | 62 | - | 0.706408 AGGATCCCACTGGGCCC |
| chr9 | 122800776 | 122800925 | + | 122800802 | 122800818 | -49 | -34 | + | 0.70631 GGCAGCATGCTCTTCCT |
| chr7 | 4678576 | 4678725 | + | 4678709 | 4678725 | 58 | 73 | - | 0.706007 AGGAATATAAGGTCCAC |
| chr10 | 4445976 | 4446125 | + | 4446009 | 4446025 | -42 | -27 | + | 0.705995 GAGGCCAGAGTATCCTC |
| chr7 | 1686926 | 1687075 | + | 1686990 | 1687006 | -11 | 4 | - | 0.705945 TTGGACGGCTGTGGCC |
| chr15 | 100016226 | 100016375 | + | 100016303 | 100016319 | 2 | 17 | - | 0.705671 GGGAACATCTTCCTTCT |
| chr10 | 103326626 | 103326775 | + | 103326724 | 103326740 | 23 | 38 | - | 0.705632 AGGCACAACCCTGACCT |
| chr1 | 1097226 | 1097375 | + | 1097349 | 1097365 | 48 | 63 | - | 0.705624 AAGCAAACACTTTGCCA |
| chr12 | 49366076 | 49366225 | + | 49366080 | 49366096 | -71 | -56 | - | 0.705479 GGGGACAGCTTTGGCCT |
| chr1 | 1957676 | 1957825 | + | 1957766 | 1957782 | 15 | 30 | - | 0.705206 AGGGGCGCCTTCTTCTT |
| chr14 | 106229476 | 106229625 | + | 106229559 | 106229575 | 8 | 23 | - | 0.704909 GAGAAAGCTCCGTGCAC |
| chr17 | 79109526 | 79109675 | + | 79109556 | 79109572 | -45 | -30 | + | 0.704829 GTCTCCAGGTTGTTCCC |
| chr5 | 74532226 | 74532375 | + | 74532299 | 74532315 | -2 | 13 | + | 0.704054 AAGCACATGTTCTGGTC |
| chr21 | 46973126 | 46973275 | + | 46973134 | 46973150 | -67 | -52 | - | 0.703812 AGGCACCGCGCATCCCT |
| chr5 | 140710276 | 140710425 | + | 140710335 | 140710351 | -16 | -1 | + | 0.703637 GGGAATATTCACTACTC |
| chr4 | 187071076 | 187071225 | + | 187071199 | 187071215 | 48 | 63 | + | 0.703198 AAGTAAAAGCTGTGGCA |
| chr14 | 94451326 | 94451475 | + | 94451407 | 94451423 | 6 | 21 | + | 0.702249 GAGAACTGAGTGGATTT |
| chr12 | 44857976 | 44858125 | + | 44857988 | 44858004 | -63 | -48 | + | 0.702243 AGGATCAAAATGTAAAA |
| chr12 | 44857926 | 44858075 | + | 44857988 | 44858004 | -13 | 2 | + | 0.702243 AGGATCAAAATGTAAAA |
| chr7 | 158800526 | 158800675 | + | 158800654 | 158800670 | 53 | 68 | + | 0.702034 GAAAAGAGGGTGGACCA |
| chr15 | 102094576 | 102094725 | + | 102094683 | 102094699 | 32 | 47 | - | 0.702032 TGTTAGGTTGCGTACTT |
| chr19 | 3820976 | 3821125 | + | 3821076 | 3821092 | 25 | 40 | - | 0.702014 TGGCCGGCAGTGTCCCT |
| chr3 | 129326676 | 129326825 | + | 129326757 | 129326773 | 6 | 21 | - | 0.701989 TGCCAGACACTGTGCCA |
| chr3 | 129326626 | 129326775 | + | 129326757 | 129326773 | 56 | 71 | - | 0.701989 TGCCAGACACTGTGCCA |
| chr22 | 40051426 | 40051575 | + | 40051538 | 40051554 | 37 | 52 | - | 0.701972 CCGTACCTATGGTTCCC |
| chr7 | 30717926 | 30718075 | + | 30718047 | 30718063 | 46 | 61 | + | 0.701661 AGGAGACTCGTGTGCTT |
| chr8 | 94508376 | 94508525 | + | 94508413 | 94508429 | -38 | -23 | + | 0.701616 GGGCAAGCAGAGCACCC |
| chr6 | 168629826 | 168629975 | + | 168629886 | 168629902 | -15 | 0 | - | 0.701485 AGGCTCACCCGCTTCTT |
| chr6 | 168629776 | 168629925 | + | 168629886 | 168629902 | 35 | 50 | - | 0.701485 AGGCTCACCCGCTTCTT |
| chr14 | 106174276 | 106174425 | + | 106174345 | 106174361 | -6 | 9 | - | 0.70148 GTGCACGTGAGGTCGC |
| chr1 | 7601926 | 7602075 | + | 7602020 | 7602036 | 19 | 34 | + | 0.701279 CAGAACCAAGCGTTACA |
| chr1 | 1974726 | 1974875 | + | 1974817 | 1974833 | 16 | 31 | - | 0.701166 TGGCAGGGCGTGTCCGC |
| chr1 | 1974676 | 1974825 | + | 1974817 | 1974833 | 66 | 81 | - | 0.701166 TGGCAGGGCGTGTCCGC |
| chr17 | 14206926 | 14207075 | + | 14206996 | 14207012 | -5 | 10 | - | 0.701148 AGGCACGTGAGGGCCTC |
| chr17 | 14206876 | 14207025 | + | 14206996 | 14207012 | 45 | 60 | - | 0.701148 AGGCACGTGAGGGCCTC |
| chr6 | 168617376 | 168617525 | + | 168617434 | 168617450 | -17 | -2 | + | 0.701095 GAGCCCAGACTGCTCTG |
| chr6 | 168617326 | 168617475 | + | 168617434 | 168617450 | 33 | 48 | + | 0.701095 GAGCCCAGACTGCTCTG |
| chr15 | 75019226 | 75019375 | + | 75019256 | 75019272 | -45 | -30 | - | 0.70102 TAGCTCAACCTGGCCCC |
| chr16 | 1316076 | 1316225 | + | 1316101 | 1316117 | -50 | -35 | - | 0.700756 GAGGACCCCTCGAGCCC |
| chr17 | 60214576 | 60214725 | + | 60214664 | 60214680 | 13 | 28 | + | 0.700195 GGGAAGCGGTTCTTCTG |
| chr17 | 60214526 | 60214675 | + | 60214664 | 60214680 | 63 | 78 | + | 0.700195 GGGAAGCCGTTCTTCTG |
| chr15 | 66543826 | 66543975 | + | 66543881 | 66543897 | -20 | -5 | - | 0.699893 CAGCACAGCACCTAACC |
| chr1 | 7539026 | 7539175 | + | 7539035 | 7539051 | -66 | -51 | + | 0.699635 GGGCAAAGAGTGAAATC |
| chr8 | 1923026 | 1923175 | + | 1923109 | 1923125 | 8 | 23 | - | 0.699604 AAGTTCGCCCGATTCCC |

Supplementary Table. 5.2.3. (5/10)
Functional enrichment of principal component three top 1000 segments

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| chr8 | 1922976 | 1923125 | + | 1923109 | 1923125 | 58 | 73 | - | 0.699604 AAGTTCGCCCGATTCCC |
| chr17 | 34995976 | 34996125 | + | 34996056 | 34996072 | 5 | 20 | - | 0.699588 GAGAGCAGCCGGCTCCC |
| chr8 | 22018376 | 22018525 | + | 22018453 | 22018469 | 2 | 17 | + | 0.699487 GGGTAGAGTGTAGACCA |
| chr15 | 102094526 | 102094675 | + | 102094628 | 102094644 | 27 | 42 | + | 0.699484 TGGAACCGGACGTGGTT |
| chr16 | 30034726 | 30034875 | + | 30034769 | 30034785 | -32 | -17 | + | 0.699298 GAGCATAAAGTGTGATG |
| chr22 | 37215826 | 37215975 | + | 37215862 | 37215878 | -39 | -24 | + | 0.698587 AGGAAAAGAATGAAACA |
| chr10 | 2978726 | 2978875 | + | 2978767 | 2978783 | -34 | -19 | + | 0.698239 TCGTATGAAGTGTGCTT |
| chr1 | 38517126 | 38517275 | + | 38517155 | 38517171 | -46 | -31 | + | 0.698002 GGCTCCAGGATGCTCCA |
| chr11 | 72973976 | 72974125 | + | 72973987 | 72974003 | -64 | -49 | + | 0.697991 GGGAAGAGTGTGAAGTG |
| chr7 | 3488726 | 3488875 | + | 3488745 | 3488761 | -56 | -41 | - | 0.697775 GGATTCATTGAGTCCTT |
| chr5 | 140710426 | 140710575 | + | 140710544 | 140710560 | 43 | 58 | + | 0.697644 GAGCATGCCGTGTCTCG |
| chr3 | 139258226 | 139258375 | + | 139258348 | 139258364 | 47 | 62 | - | 0.697486 CAGTACCCAGTGAAGTC |
| chr14 | 105052376 | 105052525 | + | 105052478 | 105052494 | 27 | 42 | - | 0.697426 TGGAATATTATGCAGCC |
| chr3 | 66139526 | 66139675 | + | 66139640 | 66139656 | 39 | 54 | + | 0.697327 AGGAAAATAAGATTCCG |
| chr11 | 4842976 | 4843125 | + | 4843047 | 4843063 | -4 | 11 | - | 0.697275 AGATACACTACCATCCT |
| chr16 | 85201726 | 85201875 | + | 85201794 | 85201810 | -7 | 8 | + | 0.696706 TGGAGCCCGACGTCCTG |
| chr15 | 78186426 | 78186575 | + | 78186535 | 78186551 | 34 | 49 | + | 0.696679 GAGCACAGCAGCTGCTG |
| chr15 | 29611876 | 29612025 | + | 29611903 | 29611919 | -48 | -33 | - | 0.696666 CAGAACAAACTGCCACT |
| chr8 | 143546776 | 143546925 | + | 143546876 | 143546892 | 25 | 40 | - | 0.696452 GGGTTGAGTATGCTCTC |
| chr7 | 73790676 | 73790825 | + | 73790750 | 73790766 | -1 | 14 | - | 0.696219 GAGACCGCCATGCACGT |
| chr2 | 242151476 | 242151625 | + | 242151594 | 242151610 | 43 | 58 | - | 0.695944 GTCTACATTGCCTTCTT |
| chr21 | 44724626 | 44724775 | + | 44724649 | 44724665 | -52 | -37 | - | 0.695935 GGCCACACCGTTGCCCC |
| chr7 | 5319476 | 5319625 | + | 5319610 | 5319626 | 59 | 74 | - | 0.695828 CCGAAATTACTGTACCT |
| chr19 | 45003726 | 45003875 | + | 45003825 | 45003841 | 24 | 39 | + | 0.69574 AAGCGCGTTGTCTGCTC |
| chr5 | 2205876 | 2206025 | + | 2205908 | 2205924 | -43 | -28 | - | 0.695739 CTGTACACACTGCGGCC |
| chr5 | 2205826 | 2205975 | + | 2205908 | 2205924 | 7 | 22 | - | 0.695739 CTGTACACACTGCGGCC |
| chr1 | 6531076 | 6531225 | + | 6531189 | 6531205 | 38 | 53 | - | 0.695638 GGTTACCGCGGCGCTCCC |
| chr7 | 158818076 | 158818225 | + | 158818153 | 158818169 | 2 | 17 | - | 0.695415 GACCACCACACGTGCCC |
| chr7 | 1328926 | 1329075 | + | 1328962 | 1328978 | -39 | -24 | - | 0.695398 TAGCAGAAGCTGTGCGT |
| chr7 | 120967776 | 120967925 | + | 120967866 | 120967882 | 15 | 30 | - | 0.695041 CTAAACACGTTGCACTT |
| chr10 | 3280976 | 3281125 | + | 3281056 | 3281072 | 5 | 20 | + | 0.694725 GGGCACAGAGGGAATCG |
| chr5 | 73969076 | 73969225 | + | 73969189 | 73969205 | 38 | 53 | + | 0.694106 TAGAAAAATGTGATCAT |
| chr10 | 4358476 | 4358625 | + | 4358573 | 4358589 | 22 | 37 | - | 0.693686 AGACTCTCTTTGTTCTC |
| chr10 | 134303826 | 134303975 | + | 134303916 | 134303932 | 15 | 30 | + | 0.693092 GGCCACCAGCTGTTCAT |
| chr7 | 1684526 | 1684675 | + | 1684554 | 1684570 | -47 | -32 | - | 0.692822 GGGAGAGGCCTGTCCTC |
| chr6 | 168629726 | 168629875 | + | 168629781 | 168629797 | -20 | -5 | - | 0.692428 AATTTCGGTCTGTTCCG |
| chr10 | 3602676 | 3602825 | + | 3602712 | 3602728 | -39 | -24 | + | 0.692333 AGGTAAAGACACTACCT |
| chr10 | 3602626 | 3602775 | + | 3602712 | 3602728 | 11 | 26 | + | 0.692333 AGGTAAAGACACTACCT |
| chr21 | 41027776 | 41027925 | + | 41027875 | 41027891 | 24 | 39 | + | 0.692166 GTGGGCACAGTGGTCTG |
| chr7 | 1407276 | 1407425 | + | 1407403 | 1407419 | 52 | 67 | - | 0.692136 GATTTCATTTTGCACCT |
| chr1 | 229480026 | 229480175 | + | 229480127 | 229480143 | 26 | 41 | - | 0.691929 GGTCAGACCGTCTGCTC |
| chr16 | 474226 | 474375 | + | 474327 | 474343 | 26 | 41 | - | 0.691607 GAATGCAGTATCTTCCT |
| chr1 | 1936476 | 1936625 | + | 1936508 | 1936524 | -43 | -28 | - | 0.691354 CTGTGCATCCTGTTACA |
| chr1 | 1936426 | 1936575 | + | 1936508 | 1936524 | 7 | 22 | - | 0.691354 CTGTGCATCCTGTTACA |
| chr1 | 1936376 | 1936525 | + | 1936508 | 1936524 | 57 | 72 | - | 0.691354 CTGTGCATCCTGTTACA |
| chr5 | 1207326 | 1207475 | + | 1207371 | 1207387 | -30 | -15 | + | 0.69113 GAGCACCTTGGGCACCG |
| chr10 | 3708426 | 3708575 | + | 3708440 | 3708456 | -61 | -46 | - | 0.691057 CAAAACACACAGTTGCT |
| chr10 | 3708376 | 3708525 | + | 3708440 | 3708456 | -11 | 4 | - | 0.691057 CAAAACACACAGTTGCT |
| chr10 | 3708326 | 3708475 | + | 3708440 | 3708456 | 39 | 54 | - | 0.691057 CAAAACACACAGTTGCT |
| chr14 | 100631676 | 100631825 | + | 100631705 | 100631721 | -46 | -31 | - | 0.691 CAGACCCCAGCGTTCGC |
| chr9 | 6716276 | 6716425 | + | 6716374 | 6716390 | 23 | 38 | + | 0.690858 ACGATCGCTGGGTTCTA |
| chr6 | 33561276 | 33561425 | + | 33561278 | 33561294 | -73 | -58 | - | 0.690722 CAGATGATGATGTTGCT |
| chr20 | 61979276 | 61979425 | + | 61979283 | 61979299 | -68 | -53 | + | 0.690628 AGGCAAACGGGGTCACT |
| chr20 | 61979226 | 61979375 | + | 61979283 | 61979299 | -18 | -3 | + | 0.690628 AGGCAAACGGGGTCACT |
| chr19 | 4566426 | 4566575 | + | 4566500 | 4566516 | -1 | 14 | - | 0.6906 AAGACCATCTGCTTCTC |
| chr14 | 70348376 | 70348525 | + | 70348445 | 70348461 | -6 | 9 | + | 0.690273 GAGAAGGTGCTGTCTCG |
| chr14 | 70348326 | 70348475 | + | 70348445 | 70348461 | 44 | 59 | + | 0.690273 GAGAAGGTGCTGTCTCG |
| chr2 | 219487476 | 219487625 | + | 219487574 | 219487590 | 23 | 38 | - | 0.690107 TTGACCGCTATGAACCT |
| chr7 | 131831476 | 131831625 | + | 131831507 | 131831523 | -44 | -29 | - | 0.690057 TAGCCCATCACTTACCA |
| chr2 | 219487426 | 219487575 | + | 219487565 | 219487581 | 64 | 79 | - | 0.689272 TGGAACTCATTGACCGC |
| chr12 | 116944026 | 116944175 | + | 116944163 | 116944179 | 62 | 77 | + | 0.68912 GGGAACAGACTGAGGAC |
| chr11 | 64739726 | 64739875 | + | 64739777 | 64739793 | -24 | -9 | - | 0.688978 CGTTATGCAGGGTACCT |
| chr11 | 64739676 | 64739825 | + | 64739777 | 64739793 | 26 | 41 | - | 0.688978 CGTTATGCAGGGTACCT |
| chr10 | 4194426 | 4194575 | + | 4194499 | 4194515 | -2 | 13 | - | 0.688912 GTGACAACCACGTCCCT |
| chr7 | 72787926 | 72788075 | + | 72787951 | 72787967 | -50 | -35 | - | 0.688623 CAGAACTACCCCTTCCG |
| chr11 | 67219426 | 67219575 | + | 67219531 | 67219547 | 30 | 45 | - | 0.688505 TGCGACAGGCTGTGGCC |
| chr1 | 1936526 | 1936675 | + | 1936610 | 1936626 | 9 | 24 | - | 0.688453 CGGGCCACTCTATGCCA |
| chr7 | 1686576 | 1686725 | + | 1686582 | 1686598 | -69 | -54 | - | 0.688256 AAGTTCAGAGGGCTCCA |
| chr17 | 77644426 | 77644575 | + | 77644564 | 77644580 | 63 | 78 | + | 0.688047 CAGCACCCGGCCTGCCA |
| chr6 | 112132876 | 112133025 | + | 112132902 | 112132918 | -49 | -34 | + | 0.687475 ACCTACAAGTTGTAGCC |
| chr6 | 112132826 | 112132975 | + | 112132902 | 112132918 | 1 | 16 | + | 0.687475 ACCTACAAGTTGTAGCC |
| chr19 | 17571476 | 17571625 | + | 17571571 | 17571587 | 20 | 35 | - | 0.687453 AGCACCAGCCGGTTCTC |
| chr10 | 28971126 | 28971275 | + | 28971207 | 28971223 | 6 | 21 | + | 0.686807 GGGCACGAGAACTGCTG |
| chr11 | 2293126 | 2293275 | + | 2293164 | 2293180 | -37 | -22 | + | 0.686672 GGGAAACGCGTGCCCTC |
| chr4 | 24796526 | 24796675 | + | 24796539 | 24796555 | -62 | -47 | + | 0.686621 GAGTGCGGCTAGTGCCA |
| chr4 | 24796476 | 24796625 | + | 24796539 | 24796555 | -12 | 3 | + | 0.686621 GAGTGCGGCTAGTGCCA |
| chr5 | 131350026 | 131350175 | + | 131350094 | 131350110 | -7 | 8 | + | 0.686429 AAGAAAGTCATGCACAT |
| chr11 | 2293176 | 2293325 | + | 2293214 | 2293230 | -37 | -22 | + | 0.686315 GGTGACCAGATGCTCCT |
| chr12 | 322076 | 322225 | + | 322111 | 322127 | -40 | -25 | - | 0.686276 AAACACACCAATTTCCA |
| chr16 | 474176 | 474325 | + | 474178 | 474194 | -73 | -58 | + | 0.686071 GGGAGCCTGCTGTCTCT |
| chr16 | 474126 | 474275 | + | 474178 | 474194 | -23 | -8 | + | 0.686071 GGGAGCCTGCTGTCTCT |
| chr14 | 104639726 | 104639875 | + | 104639786 | 104639802 | -15 | 0 | + | 0.686006 TCCCACATGTTGTTCAC |
| chr1 | 17022476 | 17022625 | + | 17022597 | 17022613 | 46 | 61 | + | 0.685924 GCCAACATGGTGAAACC |
| chr14 | 100631726 | 100631875 | + | 100631766 | 100631782 | -35 | -20 | + | 0.685514 GAGAACTGTGAGTGACA |
| chr2 | 239048526 | 239048675 | + | 239048656 | 239048672 | 55 | 70 | + | 0.68505 TGGAATGCAGGGTTCAG |
| chr17 | 1811226 | 1811375 | + | 1811272 | 1811288 | -29 | -14 | + | 0.684835 AGGCAGAGCCTGCGCTG |
| chr11 | 2190026 | 2190175 | + | 2190070 | 2190086 | -31 | -16 | + | 0.684706 CAGGCCAGGGTTTGCCA |
| chr6 | 37502976 | 37503125 | + | 37503034 | 37503050 | -17 | -2 | - | 0.684464 AGGAAGGCTGGGTGTCT |
| chr17 | 66288726 | 66288875 | + | 66288841 | 66288857 | 40 | 55 | - | 0.684394 TTGCTCGCTGAGTTCTC |
| chr6 | 134350826 | 134350975 | + | 134350906 | 134350922 | 5 | 20 | - | 0.684385 TTAAACATTTTGTATTT |
| chr6 | 134350776 | 134350925 | + | 134350906 | 134350922 | 55 | 70 | - | 0.684385 TTAAACATTTTGTATTT |
| chr20 | 47278426 | 47278575 | + | 47278509 | 47278525 | 8 | 23 | - | 0.684303 AGGGAAACCAGCTGCCT |
| chr6 | 156954426 | 156954575 | + | 156954565 | 156954581 | 64 | 79 | + | 0.68425 GCGGACGATCTGAACTG |
| chr17 | 78456326 | 78456475 | + | 78456370 | 78456386 | -31 | -16 | - | 0.684232 TAAAACACCATTCTCCG |
| chr5 | 2537676 | 2537825 | + | 2537747 | 2537763 | -4 | 11 | - | 0.684051 TGTGGCTCAGTGTTCTT |
| chr7 | 3019076 | 3019225 | + | 3019173 | 3019189 | 22 | 37 | + | 0.684047 AGAGACACTAAGAGCCT |
| chr11 | 2206026 | 2206175 | + | 2206140 | 2206156 | 39 | 54 | - | 0.684016 AGGTCCGCCTGCTCCCC |
| chr8 | 136510476 | 136510625 | + | 136510585 | 136510601 | 34 | 49 | + | 0.683949 GTGGCCAGAATGATCTT |
| chr5 | 132944026 | 132944175 | + | 132944090 | 132944106 | -11 | 4 | + | 0.683858 GAGAACCTGGTGCTAAC |
| chr7 | 1358126 | 1358275 | + | 1358225 | 1358241 | 24 | 39 | - | 0.683401 ATGCACAAATGGTCCTA |
| chr16 | 22776026 | 22776175 | + | 22776106 | 22776122 | 5 | 20 | - | 0.683386 TTGTCCAGGATGAGCCC |

Supplementary Table. 5.2.3. (6/10)
Functional enrichment of principal component three top 1000 segments

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| chr16 | 22775976 | 22776125 | + | 22776106 | 22776122 | 55 | 70 | - | 0.683386 TTGTCCAGGATGAGCCC |
| chr7 | 44279526 | 44279675 | + | 44279664 | 44279680 | 63 | 78 | + | 0.683186 TCGCACACGGAGTGTCG |
| chr14 | 104770726 | 104770875 | + | 104770766 | 104770782 | -35 | -20 | + | 0.683159 AGAAACCCTCTATTCTG |
| chr3 | 186170626 | 186170775 | + | 186170627 | 186170643 | -74 | -59 | - | 0.682855 CGACACTCAGTCTCCCC |
| chr15 | 75019276 | 75019425 | + | 75019282 | 75019298 | -69 | -54 | - | 0.682837 CGTTACGCGCCATCCCC |
| chr1 | 203525526 | 203525675 | + | 203525609 | 203525625 | 8 | 23 | + | 0.682721 GAGAACACATTTCATTT |
| chr16 | 281326 | 281475 | + | 281436 | 281452 | 35 | 50 | + | 0.682609 AAGAAGGCTATGATTCT |
| chr10 | 1602426 | 1602575 | + | 1602481 | 1602497 | -20 | -5 | - | 0.68258 AGAAACATTATTTGATC |
| chr17 | 75789476 | 75789625 | + | 75789553 | 75789569 | 2 | 17 | + | 0.682472 GGGAATGTGCTCTACAA |
| chr8 | 143545076 | 143545225 | + | 143545099 | 143545115 | -52 | -37 | - | 0.682434 AGGTACTTTGTTCCCCA |
| chr5 | 1855976 | 1856125 | + | 1856111 | 1856127 | 60 | 75 | + | 0.681871 TGCAATACAATTTACCA |
| chr5 | 2335526 | 2335675 | + | 2335659 | 2335675 | 58 | 73 | - | 0.68162 TGCTATGCATTGTGCCA |
| chr3 | 13058826 | 13058975 | + | 13058862 | 13058878 | -39 | -24 | - | 0.681443 GCAAACAATAGGTGCTC |
| chr3 | 13058776 | 13058925 | + | 13058862 | 13058878 | 11 | 26 | - | 0.681443 GCAAACAATAGGTGCTC |
| chr11 | 67619726 | 67619875 | + | 67619738 | 67619754 | -63 | -48 | - | 0.68131 GTGAAGGGCTGGTACCC |
| chr10 | 3544576 | 3544725 | + | 3544585 | 3544601 | -66 | -51 | - | 0.681235 GTGCTCACTGGGTGCTG |
| chr10 | 3544526 | 3544675 | + | 3544585 | 3544601 | -16 | -1 | - | 0.681235 GTGCTCACTGGGTGCTG |
| chr1 | 57718876 | 57719025 | + | 57718902 | 57718918 | -49 | -34 | - | 0.681003 AGGGTAGCACTGTCCTC |
| chr22 | 37500626 | 37500775 | + | 37500685 | 37500701 | -16 | -1 | - | 0.680882 ATAAATACCGTGTAGCT |
| chr22 | 37500576 | 37500725 | + | 37500685 | 37500701 | 34 | 49 | - | 0.680882 ATAAATACCGTGTAGCT |
| chr22 | 37771126 | 37771275 | + | 37771131 | 37771147 | -70 | -55 | + | 0.680871 GAGAAGGCGGTGGGGGT |
| chr5 | 173097476 | 173097625 | + | 173097518 | 173097534 | -33 | -18 | - | 0.680813 TAAAAAAACATGAACCC |
| chr5 | 173097426 | 173097575 | + | 173097518 | 173097534 | 17 | 32 | - | 0.680813 TAAAAAAACATGAACCC |
| chr10 | 4386726 | 4386875 | + | 4386865 | 4386881 | 64 | 79 | + | 0.680799 TGGAAAACATTCTTCAG |
| chr8 | 143570826 | 143570975 | + | 143570833 | 143570849 | -68 | -53 | - | 0.680575 CAGAAACGCGTGTGCGC |
| chr14 | 106438026 | 106438175 | + | 106438119 | 106438135 | 18 | 33 | + | 0.680331 GAGGAGGCATAGTGCCA |
| chr12 | 116008026 | 116008175 | + | 116008084 | 116008100 | -17 | -2 | + | 0.680142 TAGGACAATTAGCACCA |
| chr12 | 116007976 | 116008125 | + | 116008084 | 116008100 | 33 | 48 | + | 0.680142 TAGGACAATTAGCACCA |
| chr8 | 1922926 | 1923075 | + | 1922954 | 1922970 | -47 | -32 | - | 0.679927 TACAAAACCACGTTTCC |
| chr2 | 3697426 | 3697575 | + | 3697467 | 3697483 | -34 | -19 | - | 0.6798 TGGGATTCAGTGTTTTT |
| chr7 | 2728826 | 2728975 | + | 2728895 | 2728911 | -6 | 9 | + | 0.679752 GGGGAGCATGAGTGCCC |
| chr7 | 2728776 | 2728925 | + | 2728895 | 2728911 | 44 | 59 | + | 0.679752 GGGGAGCATGAGTGCCC |
| chr15 | 33437226 | 33437375 | + | 33437313 | 33437329 | 12 | 27 | + | 0.679658 AGAAATCATGTGTGCTC |
| chr19 | 15585376 | 15585525 | + | 15585503 | 15585519 | 52 | 67 | - | 0.679548 AATGACATAATGTTGGC |
| chr7 | 99987526 | 99987675 | + | 99987601 | 99987617 | 0 | 15 | + | 0.679427 GGGCAAACGACGCTCAG |
| chr7 | 99987476 | 99987625 | + | 99987601 | 99987617 | 50 | 65 | + | 0.679427 GGGCAAACGACGCTCAG |
| chr2 | 239695676 | 239695825 | + | 239695686 | 239695702 | -65 | -50 | - | 0.679296 CACCGCACACTGTGCCG |
| chr19 | 3030226 | 3030375 | + | 3030257 | 3030273 | -44 | -29 | + | 0.679205 CAGAACTGGAAGAACCA |
| chr19 | 3030176 | 3030325 | + | 3030257 | 3030273 | 6 | 21 | + | 0.679205 CAGAACTGGAAGAACCA |
| chr1 | 22975576 | 22975725 | + | 22975606 | 22975622 | -45 | -30 | + | 0.678915 GAGAACCTGGCCTGTCC |
| chr1 | 22975526 | 22975675 | + | 22975606 | 22975622 | 5 | 20 | + | 0.678915 GAGAACCTGGCCTGTCC |
| chr21 | 46799776 | 46799925 | + | 46799802 | 46799818 | -49 | -34 | - | 0.678868 CAGAACAGAAGCTCCTG |
| chr9 | 139925726 | 139925875 | + | 139925734 | 139925750 | -67 | -52 | + | 0.678573 GGGCACAAAGATGTCCG |
| chr8 | 144212926 | 144213075 | + | 144212976 | 144212992 | -25 | -10 | + | 0.678228 AAGACCGTGATGGAACC |
| chr7 | 73465926 | 73466075 | + | 73465953 | 73465969 | -48 | -33 | - | 0.678202 GGAGACACCCGGGCCCC |
| chr7 | 73465876 | 73466025 | + | 73465953 | 73465969 | 2 | 17 | - | 0.678202 GGAGACACCCGGGCCCC |
| chr17 | 25798626 | 25798775 | + | 25798691 | 25798707 | -10 | 5 | + | 0.678137 AGCAAACAGGTGTCCTC |
| chr11 | 120044426 | 120044575 | + | 120044512 | 120044528 | 11 | 26 | + | 0.678075 GGCTACACAGCTAACTG |
| chr7 | 95154926 | 95155075 | + | 95154932 | 95154948 | -69 | -54 | - | 0.678028 AATTTCATCCTGACCCT |
| chr4 | 8158176 | 8158325 | + | 8158217 | 8158233 | -34 | -19 | - | 0.677961 GGCAACTGCATGGCCTC |
| chr1 | 172291626 | 172291775 | + | 172291667 | 172291683 | -34 | -19 | - | 0.677959 AGGAGAGCCATTTGCCC |
| chr1 | 172291576 | 172291725 | + | 172291667 | 172291683 | 16 | 31 | - | 0.677959 AGGAGAGCCATTTGCCC |
| chr14 | 70037526 | 70037675 | + | 70037545 | 70037561 | -56 | -41 | - | 0.677541 TTGAACATTATCTGCAG |
| chr5 | 140710376 | 140710525 | + | 140710438 | 140710454 | -13 | 2 | - | 0.677265 GAGTCCGCATCGTCTCC |
| chr2 | 27938026 | 27938175 | + | 27938093 | 27938109 | -8 | 7 | + | 0.67726 GGGGACAGGATGCAAGG |
| chr9 | 34588426 | 34588575 | + | 34588504 | 34588520 | 3 | 18 | - | 0.677253 CAGAACAACTAGTGTTA |
| chr14 | 101123376 | 101123525 | + | 101123382 | 101123398 | -69 | -54 | + | 0.67721 ATGCACCCGGTTTCCTG |
| chr8 | 143535526 | 143535675 | + | 143535527 | 143535543 | -74 | -59 | + | 0.67715 GTGCACAGCCTGTTGAG |
| chr8 | 143535426 | 143535575 | + | 143535527 | 143535543 | 26 | 41 | + | 0.67715 GTGCACAGCCTGTTGAG |
| chr3 | 139258176 | 139258325 | + | 139258312 | 139258328 | 61 | 76 | - | 0.676829 AGGTACTCCTCGAAATT |
| chr16 | 90115026 | 90115175 | + | 90115118 | 90115134 | 17 | 32 | + | 0.676344 GGCAACGTATTTTCCTA |
| chr16 | 876176 | 876325 | + | 876245 | 876261 | -6 | 9 | - | 0.676249 TGGTCGCCACTGTGCCT |
| chr16 | 876126 | 876275 | + | 876245 | 876261 | 44 | 59 | - | 0.676249 TGGTCGCCACTGTGCCT |
| chr17 | 78981976 | 78982125 | + | 78982021 | 78982037 | -30 | -15 | + | 0.675992 AGGAACAGGACCCTTCC |
| chr17 | 78981926 | 78982075 | + | 78982021 | 78982037 | 20 | 35 | + | 0.675992 AGGAACAGGACCCTTCC |
| chr4 | 3288326 | 3288475 | + | 3288455 | 3288471 | 54 | 69 | + | 0.675962 AGCGGGAACGTGTTCCC |
| chr1 | 22889176 | 22889325 | + | 22889262 | 22889278 | 11 | 26 | - | 0.675927 GAGTTCCACCTGGGCCC |
| chr1 | 1920726 | 1920875 | + | 1920767 | 1920783 | -34 | -19 | + | 0.675885 CAGGACAGGGCGGGACGG |
| chr1 | 1920676 | 1920825 | + | 1920767 | 1920783 | 16 | 31 | + | 0.675885 CAGGACAGGGCGGGACGG |
| chr1 | 1920626 | 1920775 | + | 1920767 | 1920783 | 66 | 81 | + | 0.675885 CAGGACAGGGCGGACGG |
| chr10 | 130959526 | 130959675 | + | 130959550 | 130959566 | -51 | -36 | - | 0.675819 TGGAATGTCCCTTCCCC |
| chr6 | 46455876 | 46456025 | + | 46455959 | 46455975 | 8 | 23 | + | 0.675775 AGGTGCACGTCTTACAA |
| chr6 | 46455826 | 46455975 | + | 46455959 | 46455975 | 58 | 73 | + | 0.675775 AGGTGCACGTCTTACAA |
| chr5 | 111089976 | 111090125 | + | 111089992 | 111090008 | -59 | -44 | + | 0.675766 AGAAACATGGTATCTTA |
| chr19 | 44146826 | 44146975 | + | 44146860 | 44146876 | -41 | -26 | - | 0.675658 AATAACACCGTTTTTGC |
| chr7 | 1686676 | 1686825 | + | 1686730 | 1686746 | -21 | -6 | - | 0.675447 CTGCACGGGCTGACCCC |
| chr7 | 1686626 | 1686775 | + | 1686730 | 1686746 | 29 | 44 | - | 0.675447 CTGCACGGGCTGACCCC |
| chr1 | 61407976 | 61408125 | + | 61408020 | 61408036 | -31 | -16 | - | 0.675395 TGGTGCAATTGTTTCCC |
| chr2 | 159705326 | 159705475 | + | 159705392 | 159705408 | -9 | 6 | + | 0.675326 AAGGTCCAGCCGTTCCA |
| chr2 | 159705276 | 159705425 | + | 159705392 | 159705408 | 41 | 56 | + | 0.675326 AAGGTCCAGCCGTTCCA |
| chr5 | 141993126 | 141993275 | + | 141993133 | 141993149 | -68 | -53 | - | 0.675216 CTGTGCAGGATTTACCA |
| chr10 | 5406926 | 5407075 | + | 5406970 | 5406986 | -31 | -16 | - | 0.67497 CGATCCACCCTGTGATC |
| chr2 | 75136476 | 75136625 | + | 75136495 | 75136511 | -56 | -41 | + | 0.674853 GCGACCGCGCGGGGTTCTG |
| chr15 | 78114776 | 78114925 | + | 78114837 | 78114853 | -14 | 1 | - | 0.674674 GGGGTCCCCTTGTGTCC |
| chr2 | 189191626 | 189191775 | + | 189191632 | 189191648 | -69 | -54 | + | 0.674577 GCAGACAGTCTGTACAA |
| chr2 | 189191576 | 189191725 | + | 189191632 | 189191648 | -19 | -4 | + | 0.674577 GCAGACAGTCTGTACAA |
| chr2 | 189191526 | 189191675 | + | 189191632 | 189191648 | 31 | 46 | + | 0.674577 GCAGACAGTCTGTACAA |
| chr15 | 62543076 | 62543225 | + | 62543089 | 62543105 | -62 | -47 | - | 0.674494 GAGTGAGCCTTTTTCCC |
| chr15 | 62543026 | 62543175 | + | 62543089 | 62543105 | -12 | 3 | - | 0.674494 GAGTGAGCCTTTTTCCC |
| chr17 | 41739176 | 41739325 | + | 41739249 | 41739265 | -2 | 13 | - | 0.674353 GAGTTCGTCCTGGAGCC |
| chr17 | 41739126 | 41739275 | + | 41739249 | 41739265 | 48 | 63 | - | 0.674353 GAGTTCGTCCTGGAGCC |
| chr5 | 497426 | 497575 | + | 497565 | 497581 | 64 | 79 | + | 0.674148 GGGCACCGGGGACTGCTG |
| chr12 | 52238926 | 52239075 | + | 52239009 | 52239025 | 8 | 23 | - | 0.674123 AAGTACACTCCCCATCC |
| chr12 | 52238876 | 52239025 | + | 52239009 | 52239025 | 58 | 73 | - | 0.674123 AAGTACACTCCCCATCC |
| chr12 | 1675826 | 1675975 | + | 1675924 | 1675940 | 23 | 38 | - | 0.674009 AGCTTCCCCGAGTTCTC |
| chr11 | 60482376 | 60482525 | + | 60482401 | 60482417 | -50 | -35 | + | 0.673979 AGCCACAAAGAGTCCAT |
| chr7 | 24328476 | 24328625 | + | 24328571 | 24328587 | 20 | 35 | + | 0.673955 AAGCAGGGGGTTGTTTTA |
| chr14 | 104627776 | 104627925 | + | 104627840 | 104627856 | -11 | 4 | - | 0.673915 AAGGAAAATATTTCCTT |
| chr14 | 103691426 | 103691575 | + | 103691529 | 103691545 | 28 | 43 | + | 0.673854 AGTTACAGTCAGAGCCC |
| chr5 | 2206976 | 2207125 | + | 2207006 | 2207022 | -45 | -30 | + | 0.673832 AAGAACACATTTCAACA |
| chr5 | 2206926 | 2207075 | + | 2207006 | 2207022 | 5 | 20 | + | 0.673832 AAGAACACATTTCAACA |

Supplementary Table. 5.2.3. (7/10)
Functional enrichment of principal component three top 1000 segments

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| chr10 | 134610376 | 134610525 | + | 134610435 | 134610451 | -16 | -1 | - | 0.673802 | TAAACCTTTCTGTTCTT |
| chr10 | 134610326 | 134610475 | + | 134610435 | 134610451 | 34 | 49 | - | 0.673802 | TAAACCTTTCTGTTCTT |
| chr20 | 21483826 | 21483975 | + | 21483902 | 21483918 | 1 | 16 | + | 0.673781 | TGGGAAATGGCGAACAT |
| chr16 | 29267726 | 29267875 | + | 29267866 | 29267882 | 65 | 80 | + | 0.673646 | AGGAACCTGCGCACTCCT |
| chr21 | 46799826 | 46799975 | + | 46799835 | 46799851 | -66 | -51 | + | 0.673612 | GGCCTCACCGTGACCCA |
| chr1 | 154843126 | 154843275 | + | 154843214 | 154843230 | 13 | 28 | + | 0.673493 | GTGGGCAGTATGGTCTC |
| chr6 | 37014426 | 37014575 | + | 37014460 | 37014476 | -41 | -26 | + | 0.673435 | AAAGACAGGATGTTTAA |
| chr14 | 94463676 | 94463825 | + | 94463810 | 94463826 | 59 | 74 | + | 0.673321 | AGGAACAGTTTGCAAGA |
| chr1 | 226791376 | 226791525 | + | 226791503 | 226791519 | 52 | 67 | - | 0.673257 | TGTCGCGTCCTGTCCTC |
| chr19 | 47735926 | 47736075 | + | 47735928 | 47735944 | -73 | -58 | - | 0.672475 | TGGTGAGTTTTGTCCTT |
| chr19 | 47735876 | 47736025 | + | 47735928 | 47735944 | -23 | -8 | - | 0.672475 | TGGTGAGTTTTGTCCTT |
| chr5 | 2207076 | 2207225 | + | 2207204 | 2207220 | 53 | 68 | + | 0.672187 | TGGTGCCTGAAGTCCTT |
| chr1 | 38513526 | 38513675 | + | 38513565 | 38513581 | -36 | -21 | + | 0.672174 | AGGAGAAGGATGGGCTC |
| chr1 | 38513476 | 38513625 | + | 38513565 | 38513581 | 14 | 29 | + | 0.672174 | AGGAGAAGGATGGGCTC |
| chr10 | 130959626 | 130959775 | + | 130959725 | 130959741 | 24 | 39 | - | 0.671597 | CACCAGGCACTGTGCCT |
| chr10 | 3250626 | 3250775 | + | 3250719 | 3250735 | 18 | 33 | - | 0.671355 | CAGGTCTTCCTTTTCCC |
| chr14 | 101927926 | 101928075 | + | 101928046 | 101928062 | 45 | 60 | - | 0.671275 | TAGAATTTCCTGGCCCC |
| chr11 | 2293226 | 2293375 | + | 2293254 | 2293270 | -47 | -32 | + | 0.671055 | GGGAACGGAGATTTTTG |
| chr1 | 121260626 | 121260775 | + | 121260694 | 121260710 | -7 | 8 | + | 0.670996 | AGGAGGAAAGTGTCTCT |
| chr15 | 67841326 | 67841475 | + | 67841368 | 67841384 | -33 | -18 | - | 0.670847 | GGAAAGGCTGTGGGCTC |
| chr15 | 67841276 | 67841425 | + | 67841368 | 67841384 | 17 | 32 | - | 0.670847 | GGAAAGGCTGTGGGCTC |
| chr17 | 27396876 | 27397025 | + | 27396986 | 27397002 | 35 | 50 | + | 0.670678 | AGGACCTCGAGGTTCAA |
| chr20 | 4705176 | 4705325 | + | 4705299 | 4705315 | 48 | 63 | + | 0.670295 | TGGAACCGGAAGGCCCT |
| chr8 | 54164676 | 54164825 | + | 54164685 | 54164701 | -66 | -51 | - | 0.670201 | TATTTCCTTGCGTTCTC |
| chr8 | 54164626 | 54164775 | + | 54164685 | 54164701 | -16 | -1 | - | 0.670201 | TATTTCCTTGCGTTCTC |
| chr10 | 3526676 | 3526825 | + | 3526786 | 3526802 | 35 | 50 | + | 0.670174 | GAGCACACATTTATGCT |
| chr8 | 144303126 | 144303275 | + | 144303158 | 144303174 | -43 | -28 | + | 0.670135 | AGGGGCATCCAGCACTC |
| chr17 | 27396826 | 27396975 | + | 27396909 | 27396925 | 8 | 23 | - | 0.670076 | GAGCTCTCGGTTTCCCA |
| chr1 | 2527426 | 2527575 | + | 2527559 | 2527575 | 58 | 73 | + | 0.66983 | ATGGCCCTCGTGTCCCA |
| chr10 | 4482576 | 4482725 | + | 4482694 | 4482710 | 43 | 58 | + | 0.669602 | AGGTATCCCGTGATGCT |
| chr16 | 88963626 | 88963775 | + | 88963731 | 88963747 | 30 | 45 | + | 0.669477 | TAGAGCCGGGAGTGCCA |
| chr10 | 134610276 | 134610425 | + | 134610276 | 134610292 | -75 | -60 | + | 0.669223 | AGGTAACTTCTGAGCCA |
| chr10 | 134610226 | 134610375 | + | 134610276 | 134610292 | -25 | -10 | + | 0.669223 | AGGTAACTTCTGAGCCA |
| chr16 | 1316326 | 1316475 | + | 1316454 | 1316470 | 53 | 68 | - | 0.669176 | TTGAACTGCATTTGCCG |
| chr9 | 132315726 | 132315875 | + | 132315858 | 132315874 | 57 | 72 | - | 0.668873 | AAGCATATCCTTATCCT |
| chr8 | 143546726 | 143546875 | + | 143546772 | 143546788 | -29 | -14 | - | 0.668859 | GGGGACTGCAGGAGCCT |
| chr7 | 1407426 | 1407575 | + | 1407532 | 1407548 | 31 | 46 | - | 0.668768 | ATAAACACTGACTGCCC |
| chr12 | 322176 | 322325 | + | 322181 | 322197 | -70 | -55 | - | 0.668582 | AAGTTAAGTCTGTTTTC |
| chr12 | 322126 | 322275 | + | 322181 | 322197 | -20 | -5 | - | 0.668582 | AAGTTAAGTCTGTTTTC |
| chr1 | 7408726 | 7408875 | + | 7408759 | 7408775 | -42 | -27 | + | 0.668545 | AGCCACAGGGGGGTGTTC |
| chr1 | 65362376 | 65362525 | + | 65362408 | 65362424 | -43 | -28 | - | 0.668476 | CAGGACACACCCGACCA |
| chr1 | 4794826 | 4794975 | + | 4794886 | 4794902 | -15 | 0 | - | 0.668445 | GGAAGCATCTTCTGCTT |
| chr20 | 61371426 | 61371575 | + | 61371505 | 61371521 | 4 | 19 | - | 0.66813 | CAAAGCGCTCTGTGCTA |
| chr14 | 37075376 | 37075525 | + | 37075475 | 37075491 | 24 | 39 | + | 0.668023 | AGCCAAGCTGTGATCCA |
| chr7 | 3018376 | 3018525 | + | 3018397 | 3018413 | -54 | -39 | - | 0.667967 | CGGCTCACCTGTTTCCA |
| chr7 | 3018326 | 3018475 | + | 3018397 | 3018413 | -4 | 11 | - | 0.667967 | CGGCTCACCTGTTTCCA |
| chr7 | 3018276 | 3018425 | + | 3018397 | 3018413 | 46 | 61 | - | 0.667967 | CGGCTCACCTGTTTCCA |
| chr8 | 143868076 | 143868225 | + | 143868179 | 143868195 | 28 | 43 | - | 0.667854 | CTGGCCCTGCTGTTCCT |
| chr5 | 1217676 | 1217825 | + | 1217784 | 1217800 | 33 | 48 | + | 0.667783 | AAGTGCAGGATGCGCAG |
| chr14 | 65289626 | 65289775 | + | 65289649 | 65289665 | -52 | -37 | + | 0.667444 | GGCCCCAAACAGTACCT |
| chr16 | 14380676 | 14380825 | + | 14380738 | 14380754 | -13 | 2 | - | 0.667419 | GGATTCTGGGCGTCCCT |
| chr16 | 14380626 | 14380775 | + | 14380738 | 14380754 | 37 | 52 | - | 0.667419 | GGATTCTGGGCGTCCCT |
| chr16 | 1295476 | 1295625 | + | 1295574 | 1295590 | 23 | 38 | - | 0.667336 | TGTCTGACGTTGTTCTC |
| chr1 | 15685176 | 15685325 | + | 15685253 | 15685269 | 2 | 17 | - | 0.667207 | TAGGGCTCGATGTGCAT |
| chr4 | 184244676 | 184244825 | + | 184244743 | 184244759 | -8 | 7 | - | 0.666973 | AAGTAGAACCACTGCCC |
| chr7 | 3018526 | 3018675 | + | 3018635 | 3018651 | 34 | 49 | - | 0.6669 | TTGCACTTTATCTACTA |
| chr1 | 63798276 | 63798425 | + | 63798366 | 63798382 | 15 | 30 | + | 0.66658 | TGGGAACTGGTATTCTA |
| chr1 | 63798226 | 63798375 | + | 63798366 | 63798382 | 65 | 80 | + | 0.66658 | TGGGAACTGGTATTCTA |
| chr7 | 4870126 | 4870275 | + | 4870214 | 4870230 | 13 | 28 | + | 0.666302 | AAGCACGACATCAGCCA |
| chr10 | 26502026 | 26502175 | + | 26502119 | 26502135 | 18 | 33 | + | 0.666046 | GGGCTCAGAGCGTTTGT |
| chr3 | 14595676 | 14595825 | + | 14595796 | 14595812 | 45 | 60 | + | 0.665825 | ATGCAGAGCCTGATCCA |
| chr10 | 3313076 | 3313225 | + | 3313207 | 3313223 | 56 | 71 | + | 0.66569 | AGCAAGCGAGTGTTCAC |
| chr10 | 131650476 | 131650625 | + | 131650498 | 131650514 | -53 | -38 | - | 0.665596 | GAGTTAATCTCGTTACT |
| chr10 | 131650426 | 131650575 | + | 131650498 | 131650514 | -3 | 12 | - | 0.665596 | GAGTTAATCTCGTTACT |
| chr10 | 131650376 | 131650525 | + | 131650498 | 131650514 | 47 | 62 | - | 0.665596 | GAGTTAATCTCGTTACT |
| chr7 | 1388126 | 1388275 | + | 1388128 | 1388144 | -73 | -58 | - | 0.664969 | CGACACACTTCCTTCTG |
| chr7 | 1388076 | 1388225 | + | 1388128 | 1388144 | -23 | -8 | - | 0.664969 | CGACACACTTCCTTCTG |
| chr4 | 187729076 | 187729225 | + | 187729116 | 187729132 | -35 | -20 | - | 0.664943 | AGTGTCAACAAGTTCTT |
| chr4 | 187729026 | 187729175 | + | 187729116 | 187729132 | 15 | 30 | - | 0.664943 | AGTGTCAACAAGTTCTT |
| chr10 | 4296276 | 4296425 | + | 4296300 | 4296316 | -51 | -36 | - | 0.664936 | TGGAAGGGCAGGTCCTG |
| chr8 | 54164726 | 54164875 | + | 54164834 | 54164850 | 33 | 48 | - | 0.664909 | GAGGTGACAGTGTCTCT |
| chr20 | 48124076 | 48124225 | + | 48124190 | 48124206 | 39 | 54 | + | 0.664884 | AGTAACGAGGTGAATTG |
| chr1 | 1936726 | 1936875 | + | 1936796 | 1936812 | -5 | 10 | - | 0.664287 | CAGAATGCCCAGCACCC |
| chr1 | 1936676 | 1936825 | + | 1936796 | 1936812 | 45 | 60 | - | 0.664287 | CAGAATGCCCAGCACCC |
| chr16 | 3142576 | 3142725 | + | 3142650 | 3142666 | -1 | 14 | - | 0.663963 | CAGCACACTCAGGAACT |
| chr19 | 40032626 | 40032775 | + | 40032711 | 40032727 | 10 | 25 | + | 0.663763 | AATTTCCCTTTGTTCTG |
| chr11 | 57364876 | 57365025 | + | 57365007 | 57365023 | 56 | 71 | - | 0.663715 | AGTCCCATTCCGTCCCA |
| chr6 | 41435576 | 41435725 | + | 41435693 | 41435709 | 42 | 57 | - | 0.663694 | GGGTTTTCCGAGTCCCT |
| chr4 | 1504476 | 1504625 | + | 1504589 | 1504605 | 38 | 53 | + | 0.663658 | GGTGAAAGAGTCTGCCC |
| chr3 | 185420276 | 185420425 | + | 185420343 | 185420359 | -8 | 7 | + | 0.663538 | AGGTAAAAGGCATTTTT |
| chr3 | 185420226 | 185420375 | + | 185420343 | 185420359 | 42 | 57 | + | 0.663538 | AGGTAAAAGGCATTTTT |
| chr17 | 63134076 | 63134225 | + | 63134125 | 63134141 | -26 | -11 | + | 0.663338 | GAAAACGGGGTGATTTG |
| chr21 | 47398576 | 47398725 | + | 47398579 | 47398595 | -72 | -57 | + | 0.663334 | AGGAACACAAAGCAGGC |
| chr8 | 103572776 | 103572925 | + | 103572908 | 103572924 | 57 | 72 | + | 0.663052 | AAGGATGTAACATACTC |
| chr4 | 1537526 | 1537675 | + | 1537565 | 1537581 | -36 | -21 | + | 0.662842 | TAGCATTAGGTGTATCT |
| chr1 | 210501276 | 210501425 | + | 210501288 | 210501304 | -63 | -48 | + | 0.662798 | AAGAAAAGATAGCTCCA |
| chr14 | 76877876 | 76878025 | + | 76877972 | 76877988 | 21 | 36 | + | 0.662566 | GGTGGCAGCACGTGCTG |
| chr9 | 132482376 | 132482525 | + | 132482515 | 132482531 | 64 | 79 | - | 0.66252 | AGCAAGAGCCTCTCCCC |
| chr9 | 137660326 | 137660475 | + | 137660434 | 137660450 | 33 | 48 | + | 0.662418 | AAGACCCGTTTGGTCTC |
| chr9 | 137859526 | 137859675 | + | 137859605 | 137859621 | 4 | 19 | - | 0.662395 | GAGGAGGCCAGGCTCCT |
| chr9 | 137859476 | 137859625 | + | 137859605 | 137859621 | 54 | 69 | - | 0.662395 | GAGGAGGCCAGGCTCCT |
| chr20 | 23969976 | 23970125 | + | 23970002 | 23970018 | -49 | -34 | + | 0.662332 | GGATGCACGGTCTGCAT |
| chr8 | 81963276 | 81963425 | + | 81963413 | 81963429 | 62 | 77 | - | 0.662142 | CTGAAGGCTCTGCTCTT |
| chr5 | 2205726 | 2205875 | + | 2205729 | 2205745 | -72 | -57 | + | 0.662078 | AGGAGCCCCTGGTGCTG |
| chr5 | 2205676 | 2205825 | + | 2205729 | 2205745 | -22 | -7 | + | 0.662078 | AGGAGCCCCTGGTGCTG |
| chr5 | 2205626 | 2205775 | + | 2205729 | 2205745 | 28 | 43 | + | 0.662078 | AGGAGCCCCTGGTGCTG |
| chr10 | 131034726 | 131034875 | + | 131034857 | 131034873 | 56 | 71 | - | 0.662019 | TTTAACGCCATCTCCTT |
| chr22 | 43621676 | 43621825 | + | 43621722 | 43621738 | -29 | -14 | + | 0.661947 | AGAAACTAAATGTGTCA |
| chr7 | 158828176 | 158828325 | + | 158828311 | 158828327 | 60 | 75 | + | 0.661891 | GGGTAAACAACATGCAG |
| chr10 | 3591176 | 3591325 | + | 3591228 | 3591244 | -23 | -8 | + | 0.661706 | GTGTACTGGAAGTTGCT |
| chr10 | 3591126 | 3591275 | + | 3591228 | 3591244 | 27 | 42 | + | 0.661706 | GTGTACTGGAAGTTGCT |

Supplementary Table. 5.2.3. (8/10)
Functional enrichment of principal component three top 1000 segments

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| chr10 | 131650276 | 131650425 | + | 131650298 | 131650314 | -53 | -38 | - | 0.661638 | AATCACATTCCATGCCA |
| chr10 | 131650226 | 131650375 | + | 131650298 | 131650314 | -3 | 12 | - | 0.661638 | AATCACATTCCATGCCA |
| chr2 | 209271126 | 209271275 | + | 209271160 | 209271176 | -41 | -26 | + | 0.661556 | AGACGCAGGGCCTACCT |
| chr2 | 209271076 | 209271225 | + | 209271160 | 209271176 | 9 | 24 | + | 0.661556 | AGACGCAGGGCCTACCT |
| chr2 | 209271026 | 209271175 | + | 209271160 | 209271176 | 59 | 74 | + | 0.661556 | AGACGCAGGGCCTACCT |
| chr5 | 2206676 | 2206825 | + | 2206702 | 2206718 | -49 | -34 | + | 0.661241 | GAGCACACAGACTTTAC |
| chr15 | 27210226 | 27210375 | + | 27210351 | 27210367 | 50 | 65 | + | 0.661142 | AGAGCCATTGGGTGCTT |
| chr10 | 3479876 | 3480025 | + | 3480002 | 3480018 | 51 | 66 | - | 0.661093 | CTGTAGACCAAGTGCAC |
| chr7 | 127881476 | 127881625 | + | 127881490 | 127881506 | -61 | -46 | + | 0.66102 | ATGATAGCTTTGTACCG |
| chr7 | 1137276 | 1137425 | + | 1137289 | 1137305 | -62 | -47 | + | 0.660871 | GACTGCAGAGCCTACCT |
| chr7 | 1137226 | 1137375 | + | 1137289 | 1137305 | -12 | 3 | + | 0.660871 | GACTGCAGAGCCTACCT |
| chr11 | 63996726 | 63996875 | + | 63996730 | 63996746 | -71 | -56 | - | 0.660515 | CTGCTCACCCTGAGCCA |
| chr11 | 63996676 | 63996825 | + | 63996730 | 63996746 | -21 | -6 | - | 0.660515 | CTGCTCACCCTGAGCCA |
| chr7 | 95154976 | 95155125 | + | 95155009 | 95155025 | -42 | -27 | + | 0.66049 | GGGAAGACATTCTCATT |
| chr17 | 79459326 | 79459475 | + | 79459447 | 79459463 | 46 | 61 | - | 0.660466 | GGCTTCCCTGGGTTCCT |
| chr17 | 1202176 | 1202325 | + | 1202303 | 1202319 | 52 | 67 | + | 0.660395 | GGTCCCAGGATGTCTCC |
| chr4 | 6575726 | 6575875 | + | 6575788 | 6575804 | -13 | 2 | + | 0.660313 | GTGGACAGGGCTGACTC |
| chr3 | 23782776 | 23782925 | + | 23782915 | 23782931 | 64 | 79 | - | 0.660204 | TGGTGCCCTTCTTACCT |
| chr1 | 33391376 | 33391525 | + | 33391514 | 33391530 | 63 | 78 | - | 0.659986 | CAGGACTCAGTCCTCCA |
| chr15 | 27210176 | 27210325 | + | 27210314 | 27210330 | 63 | 78 | + | 0.65991 | GGGTCTACGGTGTTGGA |
| chr10 | 94448426 | 94448575 | + | 94448461 | 94448477 | -40 | -25 | - | 0.65987 | AGGGACTACATCTGTCC |
| chr17 | 1811276 | 1811425 | + | 1811333 | 1811349 | -18 | -3 | + | 0.659467 | AACAGCAGAGTGAGCCG |
| chr1 | 87994626 | 87994775 | + | 87994641 | 87994657 | -60 | -45 | + | 0.659457 | GAGAGCTGATTGGTCCA |
| chr1 | 87994576 | 87994725 | + | 87994641 | 87994657 | -10 | 5 | + | 0.659457 | GAGAGCTGATTGGTCCA |
| chr19 | 45720026 | 45720175 | + | 45720034 | 45720050 | -67 | -52 | - | 0.659282 | AAGCTCCCTGCATTCTT |
| chr14 | 103691376 | 103691525 | + | 103691405 | 103691421 | -46 | -31 | + | 0.659086 | AGGGTCCTTGCGTATTT |
| chr12 | 107297226 | 107297375 | + | 107297320 | 107297336 | 19 | 34 | + | 0.659008 | TCGTCCGGTGTGTTTTC |
| chr9 | 138109176 | 138109325 | + | 138109287 | 138109303 | 36 | 51 | - | 0.658965 | GAAAATACTTTCTGCTT |
| chr19 | 18902626 | 18902775 | + | 18902726 | 18902742 | 25 | 40 | - | 0.658838 | CGGGAAACAGAGTCACC |
| chr9 | 136500076 | 136500225 | + | 136500145 | 136500161 | -6 | 9 | - | 0.658712 | GAGAGCACTGCCTGCGA |
| chr10 | 135054876 | 135055025 | + | 135054908 | 135054924 | -43 | -28 | + | 0.658474 | AAGGTCTGGAAGTGCCC |
| chr10 | 135054826 | 135054975 | + | 135054908 | 135054924 | 7 | 22 | + | 0.658474 | AAGGTCTGGAAGTGCCC |
| chr7 | 1251076 | 1251225 | + | 1251155 | 1251171 | 4 | 19 | - | 0.658473 | CAGCAGATCCTGGTGCC |
| chr1 | 14220226 | 14220375 | + | 14220263 | 14220279 | -38 | -23 | - | 0.658206 | TCGTGTGTCATGTGCCC |
| chr1 | 14220176 | 14220325 | + | 14220263 | 14220279 | 12 | 27 | - | 0.658206 | TCGTGTGTCATGTGCCC |
| chr1 | 14220126 | 14220275 | + | 14220263 | 14220279 | 62 | 77 | - | 0.658206 | TCGTGTGTCATGTGCCC |
| chr22 | 43805226 | 43805375 | + | 43805315 | 43805331 | 14 | 29 | - | 0.658016 | TGGAAGGTGAGTTTCTT |
| chr22 | 43805176 | 43805325 | + | 43805315 | 43805331 | 64 | 79 | - | 0.658016 | TGGAAGGTGAGTTTCTT |
| chr6 | 45500826 | 45500975 | + | 45500927 | 45500943 | 26 | 41 | + | 0.657533 | AGTGAAAGGAAGGTTCTA |
| chr1 | 38605976 | 38606125 | + | 38606104 | 38606120 | 53 | 68 | - | 0.657449 | AACATCTCCATGTGCAC |
| chr3 | 72704526 | 72704675 | + | 72704568 | 72704584 | -33 | -18 | + | 0.657378 | TAGCACGTGGGCTCCTG |
| chr16 | 88880726 | 88880875 | + | 88880791 | 88880807 | -10 | 5 | + | 0.657309 | AGGCACTTGCAGGGCCA |
| chr1 | 156831076 | 156831225 | + | 156831189 | 156831205 | 38 | 53 | - | 0.657292 | GAAATCTCCGTGGACCA |
| chr10 | 4331726 | 4331875 | + | 4331734 | 4331750 | -67 | -52 | - | 0.656636 | AGGATCATTGGATAACT |
| chr1 | 7569226 | 7569375 | + | 7569258 | 7569274 | -43 | -28 | + | 0.656617 | AAGGAGAGTGTGTTGGA |
| chr19 | 46651126 | 46651275 | + | 46651132 | 46651148 | -69 | -54 | + | 0.656194 | ATGAATATGAAGTTTCA |
| chr11 | 71010376 | 71010525 | + | 71010418 | 71010434 | -33 | -18 | + | 0.656168 | AAGCGTGGCTTGTTCTG |
| chr17 | 79109626 | 79109775 | + | 79109683 | 79109699 | -18 | -3 | + | 0.65596 | GGAAACCAGCTCTGCCG |
| chr15 | 99088026 | 99088175 | + | 99088149 | 99088165 | 48 | 63 | - | 0.655604 | CAGACCCCCGTCTACAC |
| chr3 | 169540026 | 169540175 | + | 169540072 | 169540088 | -29 | -14 | + | 0.65559 | GAGCTCCGGCTCTACCA |
| chr11 | 64780626 | 64780775 | + | 64780730 | 64780746 | 29 | 44 | - | 0.655533 | AGTTAGCTATTGTCCTC |
| chr1 | 25298626 | 25298775 | + | 25298733 | 25298749 | 32 | 47 | - | 0.655431 | CAGAGCCCTCTGAGCTC |
| chr12 | 322526 | 322675 | + | 322559 | 322575 | -42 | -27 | + | 0.655293 | GGGTACAAGCACACCCA |
| chr11 | 69706676 | 69706825 | + | 69706708 | 69706724 | -43 | -28 | + | 0.655282 | GGGTCACCAATTTTCCT |
| chr17 | 66288676 | 66288825 | + | 66288803 | 66288819 | 52 | 67 | + | 0.65489 | GAGGAGTCGGCGATCTG |
| chr15 | 41219326 | 41219475 | + | 41219380 | 41219396 | -21 | -6 | - | 0.654447 | TAGATCTCTGGGATCTC |
| chr15 | 29825276 | 29825425 | + | 29825332 | 29825348 | -19 | -4 | + | 0.654356 | GGGGAAACTGAGGCCTG |
| chr1 | 1084476 | 1084625 | + | 1084499 | 1084515 | -52 | -37 | - | 0.654182 | CAGAAGAAGCAGCACCT |
| chr1 | 1084426 | 1084575 | + | 1084499 | 1084515 | -2 | 13 | - | 0.654182 | CAGAAGAAGCAGCACCT |
| chr9 | 122800826 | 122800975 | + | 122800932 | 122800948 | 31 | 46 | - | 0.653593 | GCTTCCACCATGTCCAC |
| chr4 | 3677526 | 3677675 | + | 3677665 | 3677681 | 64 | 79 | + | 0.653577 | AGGGACAGGCCCAGCTG |
| chr10 | 3457276 | 3457425 | + | 3457362 | 3457378 | 11 | 26 | + | 0.653441 | CGGGGGGGAGTGGGCCA |
| chr16 | 57317576 | 57317725 | + | 57317602 | 57317618 | -49 | -34 | - | 0.653423 | GAGTAGGAGTCTTGCCC |
| chr5 | 1923426 | 1923575 | + | 1923459 | 1923475 | -42 | -27 | - | 0.653314 | GGCTACATTTTGGCATC |
| chr16 | 1316276 | 1316425 | + | 1316346 | 1316362 | -5 | 10 | - | 0.653224 | AGGTTCTACAAGTTGCC |
| chr15 | 23894676 | 23894825 | + | 23894704 | 23894720 | -47 | -32 | - | 0.653142 | ACTTTCACCTTTTTCCT |
| chr20 | 36037626 | 36037775 | + | 36037674 | 36037690 | -27 | -12 | - | 0.652789 | TGGTCCAGCCTCAGCCT |
| chr16 | 88963576 | 88963725 | + | 88963677 | 88963693 | 26 | 41 | + | 0.65246 | AGGACAGCCCTGTGCAA |
| chr10 | 3480026 | 3480175 | + | 3480050 | 3480066 | -51 | -36 | + | 0.652443 | CGGGACAGAAAGATTTC |
| chr8 | 42009076 | 42009225 | + | 42009212 | 42009228 | 61 | 76 | + | 0.652225 | CAGCACAGGTCGATACG |
| chr4 | 3865026 | 3865175 | + | 3865132 | 3865148 | 31 | 46 | + | 0.652217 | GGGCCAAGCTTTTTCTC |
| chr14 | 104768276 | 104768425 | + | 104768318 | 104768334 | -33 | -18 | - | 0.651851 | GGACCGGCCCTGTTCTC |
| chr10 | 4414876 | 4415025 | + | 4414945 | 4414961 | -6 | 9 | + | 0.651814 | AGGACTGTTTTGTTGCT |
| chr17 | 25583176 | 25583325 | + | 25583210 | 25583226 | -41 | -26 | - | 0.651814 | ATAATCATGCTTTTCCT |
| chr2 | 71099176 | 71099325 | + | 71099221 | 71099237 | -30 | -15 | - | 0.651202 | AAGTGCTTAATATTCAC |
| chr7 | 36013226 | 36013375 | + | 36013253 | 36013269 | -48 | -33 | + | 0.651122 | AAGAAAGGGCAGATCTT |
| chr16 | 2863776 | 2863925 | + | 2863795 | 2863811 | -56 | -41 | - | 0.650998 | AAGTACTTAGAGCAGCC |
| chr1 | 6531226 | 6531375 | + | 6531277 | 6531293 | -24 | -9 | + | 0.650711 | AGCAAGCTTATGATCCC |
| chr10 | 79270626 | 79270775 | + | 79270641 | 79270657 | -60 | -45 | + | 0.650254 | ATGGAGATGACTTTCCA |
| chr11 | 120590026 | 120590175 | + | 120590053 | 120590069 | -48 | -33 | - | 0.650245 | GGGTTAACCAGGTGGCT |
| chr11 | 120589976 | 120590125 | + | 120590053 | 120590069 | 2 | 17 | - | 0.650245 | GGGTTAACCAGGTGGCT |
| chr2 | 42077526 | 42077675 | + | 42077658 | 42077674 | 57 | 72 | - | 0.650145 | TGGATGAGGGCGGTCTT |
| chr16 | 85198476 | 85198625 | + | 85198582 | 85198598 | 31 | 46 | + | 0.650099 | AGGGAAGTCGGGTTGCA |
| chr10 | 3276976 | 3277125 | + | 3277052 | 3277068 | 1 | 16 | + | 0.649895 | AGGACCGGGATCTTTTG |
| chr8 | 1707176 | 1707325 | + | 1707253 | 1707269 | 2 | 17 | - | 0.649888 | TGCTACAGAGTCTGTTT |
| chr7 | 1407226 | 1407375 | + | 1407249 | 1407265 | -52 | -37 | + | 0.649818 | GTTTAAATGATGTTTTT |
| chr7 | 1407176 | 1407325 | + | 1407249 | 1407265 | -2 | 13 | + | 0.649818 | GTTTAAATGATGTTTTT |
| chr6 | 134350726 | 134350875 | + | 134350788 | 134350804 | -13 | 2 | - | 0.649767 | AAGTTCATAACCCACCA |
| chr1 | 9341876 | 9342025 | + | 9341894 | 9341910 | -57 | -42 | - | 0.649675 | AGGATCTCCTGCTACTT |
| chr6 | 25726926 | 25727075 | + | 25726945 | 25726961 | -56 | -41 | + | 0.649643 | CCGAACGCGGCGTTTGA |
| chr7 | 3019026 | 3019175 | + | 3019111 | 3019127 | 10 | 25 | - | 0.6496 | GGGGTCACCCGCAAATC |
| chr20 | 61162126 | 61162275 | + | 61162150 | 61162166 | -51 | -36 | + | 0.649528 | TGGAACCAAATCGCACTG |
| chr10 | 4378526 | 4378675 | + | 4378635 | 4378651 | 34 | 49 | + | 0.649358 | TGAAACTTGATGTTAAT |
| chr12 | 58736226 | 58736375 | + | 58736313 | 58736329 | 12 | 27 | - | 0.649354 | AGCGAGGTCATTTTCCC |
| chr7 | 1388176 | 1388325 | + | 1388189 | 1388205 | -62 | -47 | - | 0.648723 | AATTTTACCGTTTTCTT |
| chr1 | 88108726 | 88108875 | + | 88108864 | 88108880 | 63 | 78 | + | 0.648385 | TAGCCCAGAAAATTCCT |
| chr6 | 25727276 | 25727425 | + | 25727385 | 25727401 | 34 | 49 | - | 0.648054 | CACTACAGCAAGGGCTC |
| chr20 | 23970026 | 23970175 | + | 23970040 | 23970056 | -61 | -46 | - | 0.647554 | GAGGCCTCTCTCTCCCT |
| chr11 | 14994276 | 14994425 | + | 14994315 | 14994331 | -36 | -21 | - | 0.647446 | GGGTGGGTTCTGGGCCC |
| chr11 | 14994226 | 14994375 | + | 14994315 | 14994331 | 14 | 29 | - | 0.647446 | GGGTGGGTTCTGGGCCC |
| chr21 | 46816576 | 46816725 | + | 46816579 | 46816595 | -72 | -57 | + | 0.647184 | ATGGGCGCATTGCACCT |

Supplementary Table. 5.2.3. (9/10)
Functional enrichment of principal component three top 1000 segments

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| chr21 | 46816526 | 46816675 | + | 46816579 | 46816595 | -22 | -7 | + | 0.647184 ATGGGCGCATTGCACCT |
| chr6 | 25727226 | 25727375 | + | 25727316 | 25727332 | 15 | 30 | + | 0.64671 ATGAGCATTATGAATTC |
| chr6 | 25727176 | 25727325 | + | 25727316 | 25727332 | 65 | 80 | + | 0.64671 ATGAGCATTATGAATTC |
| chr14 | 106437976 | 106438125 | + | 106438014 | 106438030 | -37 | -22 | - | 0.646573 GTGAATGACGTGTCTTT |
| chr3 | 64224976 | 64225125 | + | 64225046 | 64225062 | -5 | 10 | + | 0.645945 GGAGACATATTCTGCAA |
| chr22 | 32750776 | 32750925 | + | 32750875 | 32750891 | 24 | 39 | + | 0.645432 AGCAGCGAGGTGAGCCA |
| chr15 | 23894726 | 23894875 | + | 23894803 | 23894819 | 2 | 17 | - | 0.645372 TAAAACAAGGGTTCCTT |
| chr6 | 18990476 | 18990625 | + | 18990539 | 18990555 | -12 | 3 | + | 0.645367 AAGGAGAAAGTGATATA |
| chr6 | 18990426 | 18990575 | + | 18990539 | 18990555 | 38 | 53 | + | 0.645367 AAGGAGAAAGTGATATA |
| chr11 | 102216826 | 102216975 | + | 102216909 | 102216925 | 8 | 23 | + | 0.645244 GGGCAAAGAGCGTCAAC |
| chr11 | 102216776 | 102216925 | + | 102216909 | 102216925 | 58 | 73 | + | 0.645244 GGGCAAAGAGCGTCAAC |
| chr1 | 7408776 | 7408925 | + | 7408871 | 7408887 | 20 | 35 | + | 0.645074 ACGTTCACTGTATTACG |
| chr12 | 107297176 | 107297325 | + | 107297290 | 107297306 | 39 | 54 | - | 0.645013 AATTCCGCCCCGGACCC |
| chr5 | 149683226 | 149683375 | + | 149683333 | 149683349 | 32 | 47 | + | 0.644832 CGGACCAGTGTTTCCAG |
| chr14 | 104688476 | 104688625 | + | 104688529 | 104688545 | -22 | -7 | - | 0.644726 GGGTTGTCTCTGTGCGT |
| chr14 | 104688426 | 104688575 | + | 104688529 | 104688545 | 28 | 43 | - | 0.644726 GGGTTGTCTCTGTGCGT |
| chr16 | 84336176 | 84336325 | + | 84336219 | 84336235 | -32 | -17 | - | 0.644474 GAGCCCAGCTCTTCCTT |
| chr9 | 129387126 | 129387275 | + | 129387244 | 129387260 | 43 | 58 | + | 0.644376 AATAAGATTATGTCACG |
| chr16 | 84336226 | 84336375 | + | 84336313 | 84336329 | 12 | 27 | - | 0.644103 AAAAGCAATTCGTTCGC |
| chr7 | 3018226 | 3018375 | + | 3018251 | 3018267 | -50 | -35 | + | 0.64396 GGGCACCGGGTGCAGGC |
| chr21 | 34350976 | 34351125 | + | 34351092 | 34351108 | 41 | 56 | - | 0.643876 GAGAAAGGCCTGGAGCC |
| chr11 | 62100726 | 62100875 | + | 62100854 | 62100870 | 53 | 68 | - | 0.643717 CAGCCGCTTGTGTGCCT |
| chr10 | 26501976 | 26502125 | + | 26502035 | 26502051 | -16 | -1 | - | 0.643475 GGGTGCCCTTGGCTCCT |
| chr16 | 49530526 | 49530675 | + | 49530618 | 49530634 | 17 | 32 | + | 0.642854 GGCTACGGCAAATGCTC |
| chr10 | 134662176 | 134662325 | + | 134662230 | 134662246 | -21 | -6 | - | 0.642764 CAGCCCACTCGGGCCCC |
| chr19 | 18902576 | 18902725 | + | 18902648 | 18902664 | -3 | 12 | - | 0.64255 GAGTGCTTGCTGTATGC |
| chr3 | 55931376 | 55931525 | + | 55931438 | 55931454 | -13 | 2 | + | 0.642216 ACAGACAGGCTCTTCTG |
| chr3 | 55931326 | 55931475 | + | 55931438 | 55931454 | 37 | 52 | + | 0.642216 ACAGACAGGCTCTTCTG |
| chr17 | 7082926 | 7083075 | + | 7082967 | 7082983 | -34 | -19 | - | 0.642192 CAGAACTATTGGTGTCT |
| chr17 | 7082876 | 7083025 | + | 7082967 | 7082983 | 16 | 31 | - | 0.642192 CAGAACTATTGGTGTCT |
| chr19 | 554776 | 554925 | + | 554842 | 554858 | -9 | 6 | + | 0.642164 GTGAACTTATTTTACAA |
| chr5 | 1010826 | 1010975 | + | 1010870 | 1010886 | -31 | -16 | + | 0.642158 AGCAGCCAGGTGCTCCT |
| chr5 | 1010776 | 1010925 | + | 1010870 | 1010886 | 19 | 34 | + | 0.642158 AGCAGCCAGGTGCTCCT |
| chr5 | 532876 | 533025 | + | 532883 | 532899 | -68 | -53 | - | 0.642064 GGGAGCCCCGCCTTGCC |
| chr5 | 532826 | 532975 | + | 532883 | 532899 | -18 | -3 | - | 0.642064 GGGAGCCCCGCCTTGCC |
| chr11 | 68781826 | 68781975 | + | 68781929 | 68781945 | 28 | 43 | - | 0.642048 TGGAACGCTTGTTTACA |
| chr5 | 2207026 | 2207175 | + | 2207153 | 2207169 | 52 | 67 | + | 0.641807 AAGAACCCTGTTTATGG |
| chr10 | 34496226 | 34496375 | + | 34496243 | 34496259 | -58 | -43 | + | 0.64133 CAGCAGCAAGTGTTTTC |
| chr11 | 69706726 | 69706875 | + | 69706755 | 69706771 | -46 | -31 | - | 0.641327 TGGACAGTAAGGTGCCT |
| chr16 | 876026 | 876175 | + | 876060 | 876076 | -41 | -26 | + | 0.641287 GGTAAGATATTTTGCTG |
| chr7 | 45188076 | 45188225 | + | 45188209 | 45188225 | 58 | 73 | - | 0.641082 GACTTCGGCATGGGCCT |
| chr3 | 72704626 | 72704775 | + | 72704664 | 72704680 | -37 | -22 | + | 0.640834 CGGACAATTTTGAACTG |
| chr3 | 72704576 | 72704725 | + | 72704664 | 72704680 | 13 | 28 | + | 0.640834 CGGACAATTTTGAACTG |
| chr11 | 1102426 | 1102575 | + | 1102502 | 1102518 | 1 | 16 | + | 0.640689 ACAGACGCACTGTATCA |
| chr10 | 3500076 | 3500225 | + | 3500085 | 3500101 | -66 | -51 | - | 0.640501 TGTCACACACTGAGATC |
| chr8 | 700076 | 700225 | + | 700117 | 700133 | -34 | -19 | + | 0.640471 GTGGAAGGGCTTTTCCT |
| chr8 | 700026 | 700175 | + | 700117 | 700133 | 16 | 31 | + | 0.640471 GTGGAAGGGCTTTTCCT |
| chr1 | 57718926 | 57719075 | + | 57718936 | 57718952 | -65 | -50 | + | 0.640099 AGTGACTGTTAGTTCTA |
| chr10 | 3378976 | 3379125 | + | 3379045 | 3379061 | -6 | 9 | + | 0.640078 GGATAAAAGCTGACCCG |
| chr5 | 493226 | 493375 | + | 493331 | 493347 | 30 | 45 | + | 0.640041 GGGATAATGAGCTTCTT |
| chr12 | 52238576 | 52238725 | + | 52238616 | 52238632 | -35 | -20 | - | 0.639402 AGGCTGGCCTAGTGCCT |
| chr12 | 52238526 | 52238675 | + | 52238616 | 52238632 | 15 | 30 | - | 0.639402 AGGCTGGCCTAGTGCCT |
| chr20 | 4705126 | 4705275 | + | 4705200 | 4705216 | -1 | 14 | + | 0.639337 AGGAAGCACCTGAGCTG |
| chr21 | 43547726 | 43547875 | + | 43547815 | 43547831 | 14 | 29 | + | 0.639049 GAGAACGGCAACTCCAA |
| chr21 | 43547676 | 43547825 | + | 43547815 | 43547831 | 64 | 79 | + | 0.639049 GAGAACGGCAACTCCAA |
| chr10 | 131650676 | 131650825 | + | 131650748 | 131650764 | -3 | 12 | + | 0.639019 AGGCACTGGGATTTTCT |
| chr10 | 131650626 | 131650775 | + | 131650748 | 131650764 | 47 | 62 | + | 0.639019 AGGCACTGGGATTTTCT |
| chr15 | 29037776 | 29037925 | + | 29037825 | 29037841 | -26 | -11 | + | 0.638952 TGAAAAATATTTTTCTG |
| chr16 | 22775926 | 22776075 | + | 22775935 | 22775951 | -66 | -51 | - | 0.638545 AAGAAAGCTACCTTGCT |
| chr13 | 28563576 | 28563725 | + | 28563659 | 28563675 | 8 | 23 | + | 0.638529 TGGACTCCAGTGATCCA |
| chr8 | 144854576 | 144854725 | + | 144854619 | 144854635 | -32 | -17 | + | 0.638481 GTGTACGCAGAGGACGG |
| chr16 | 29267676 | 29267825 | + | 29267809 | 29267825 | 58 | 73 | - | 0.638456 CAAGACATGGCGTTTGT |
| chr16 | 281276 | 281425 | + | 281395 | 281411 | 44 | 59 | - | 0.638354 AGAAAAGGCACGTTTCC |
| chr9 | 6716226 | 6716375 | + | 6716229 | 6716245 | -72 | -57 | - | 0.637996 CGTTCCTCCGTTTCCCC |
| chr11 | 120764426 | 120764575 | + | 120764566 | 120764582 | 65 | 80 | - | 0.637708 TGTTAATCTCTGCACCC |
| chr17 | 75848676 | 75848825 | + | 75848780 | 75848796 | 29 | 44 | + | 0.637621 TGGAGCACGCAGCCCTG |
| chr6 | 37014476 | 37014625 | + | 37014488 | 37014504 | -63 | -48 | + | 0.637451 AGGTAAAATAAGGTCAT |
| chr12 | 107297076 | 107297225 | + | 107297205 | 107297221 | 54 | 69 | + | 0.637419 GACCACCCAGCCTGCTC |
| chr19 | 47735626 | 47735775 | + | 47735663 | 47735679 | -38 | -23 | - | 0.637263 TAGAAGAAACTGACCAC |
| chr21 | 42219676 | 42219825 | + | 42219774 | 42219790 | 23 | 38 | + | 0.637235 AGGGCTGGAATGCACCC |
| chr9 | 132383226 | 132383375 | + | 132383319 | 132383335 | 18 | 33 | - | 0.636975 AATAGCAGCACCTGCCT |
| chr5 | 1207376 | 1207525 | + | 1207505 | 1207521 | 54 | 69 | + | 0.636696 GGGCGAGTGGAGTTCAT |
| chr1 | 210612176 | 210612325 | + | 210612291 | 210612307 | 40 | 55 | + | 0.636369 ATGGACCTCATGATCAG |
| chr5 | 141993176 | 141993325 | + | 141993205 | 141993221 | -46 | -31 | + | 0.635971 GGGGCCAGATCTTTCAC |
| chr20 | 1164876 | 1165025 | + | 1164996 | 1165012 | 45 | 60 | - | 0.635741 AGTTACACAAAGGCCGT |
| chr16 | 1198576 | 1198725 | + | 1198634 | 1198650 | -17 | -2 | - | 0.635465 TAGACGGCGCTGTCCAC |
| chr16 | 1198526 | 1198675 | + | 1198634 | 1198650 | 33 | 48 | - | 0.635465 TAGACGGCGCTGTCCAC |
| chr4 | 3690826 | 3690975 | + | 3690909 | 3690925 | 8 | 23 | + | 0.635187 GAAGAGGCGGTGTGGTC |
| chr5 | 162997826 | 162997975 | + | 162997853 | 162997869 | -48 | -33 | - | 0.635128 AACTATAAAATTTGCTT |
| chr7 | 104897076 | 104897225 | + | 104897083 | 104897099 | -68 | -53 | - | 0.63478 AATGAGCCACTGTGCTC |
| chr17 | 74581226 | 74581375 | + | 74581260 | 74581276 | -41 | -26 | + | 0.634143 GGGCGCGGGGCGGACAG |
| chr3 | 97542226 | 97542375 | + | 97542278 | 97542294 | -23 | -8 | - | 0.633344 TGTTCCTCGGTTTCCCC |
| chr15 | 101807276 | 101807425 | + | 101807394 | 101807410 | 43 | 58 | - | 0.633264 GGGTAACTCTCTTGCTT |
| chr11 | 128796326 | 128796475 | + | 128796399 | 128796415 | -2 | 13 | - | 0.633221 GGCAAAACCTTCATCCT |
| chr3 | 185866476 | 185866625 | + | 185866549 | 185866565 | -2 | 13 | - | 0.632817 GAGTTCTCGGTCTCTTC |
| chr19 | 47735676 | 47735825 | + | 47735704 | 47735720 | -47 | -32 | - | 0.632121 TGTGACAGCTTCTTGCT |
| chr5 | 173737976 | 173738125 | + | 173738067 | 173738083 | 16 | 31 | + | 0.631909 GGGGTGGGCAGGTTCCA |
| chr11 | 2210126 | 2210275 | + | 2210148 | 2210164 | -53 | -38 | - | 0.631714 GAGAAATCAGAGCCCCT |
| chr14 | 101128326 | 101128475 | + | 101128332 | 101128348 | -69 | -54 | - | 0.631607 AAAAGCACCCTAGACCT |
| chr14 | 101128276 | 101128425 | + | 101128332 | 101128348 | -19 | -4 | - | 0.631607 AAAAGCACCCTAGACCT |
| chr10 | 3572226 | 3572375 | + | 3572334 | 3572350 | 33 | 48 | + | 0.631352 TGGGAAAGTCAGCACTC |
| chr15 | 99087976 | 99088125 | + | 99088072 | 99088088 | 21 | 36 | + | 0.631206 TTGACCATGAAATACCA |

Supplementary Table. 5.2.3. (10/10)
Functional enrichment of principal component three top 1000 segments

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| chr2 | 202753026 | 202753175 | + | 202753118 | 202753134 | 17 | 32 | - | 0.630757 | GATGACTGAATTTCCCT |
| chr2 | 11294426 | 11294575 | + | 11294465 | 11294481 | -36 | -21 | + | 0.630625 | CAGCACCTGTTTTTTTT |
| chr16 | 29241926 | 29242075 | + | 29241974 | 29241990 | -27 | -12 | - | 0.630598 | TGGCACAGCCAAACCCC |
| chr17 | 65527576 | 65527725 | + | 65527714 | 65527730 | 63 | 78 | + | 0.630526 | TGAAACTAACTGCACTG |
| chr12 | 52240276 | 52240425 | + | 52240277 | 52240293 | -74 | -59 | - | 0.630461 | CATAAAAAGCTGTGGCC |
| chr12 | 52240226 | 52240375 | + | 52240277 | 52240293 | -24 | -9 | - | 0.630461 | CATAAAAAGCTGTGGCC |
| chr16 | 57317626 | 57317775 | + | 57317645 | 57317661 | -56 | -41 | - | 0.630274 | AGCGATATCTCGGTCCC |
| chr10 | 118084226 | 118084375 | + | 118084254 | 118084270 | -47 | -32 | + | 0.630166 | AAGCCCGGTTTGTAAAC |
| chr11 | 67462676 | 67462825 | + | 67462772 | 67462788 | 21 | 36 | - | 0.630134 | CAGAATATCCCGACCAC |
| chr16 | 1111076 | 1111225 | + | 1111129 | 1111145 | -22 | -7 | - | 0.63008 | CAGCACCCACTCTCACT |
| chr4 | 3677476 | 3677625 | + | 3677601 | 3677617 | 50 | 65 | - | 0.630008 | GCGGACACCCGGGAACC |
| chr17 | 25798326 | 25798475 | + | 25798327 | 25798343 | -74 | -59 | - | 0.629929 | CAGACAGACGTGCTCTC |
| chr4 | 1160676 | 1160825 | + | 1160702 | 1160718 | -49 | -34 | + | 0.629818 | GGGGACAGCCAGGGACA |
| chr6 | 25727076 | 25727225 | + | 25727098 | 25727114 | -53 | -38 | + | 0.62958 | AAAAGAACCGTGTAACG |
| chr13 | 50703376 | 50703525 | + | 50703464 | 50703480 | 13 | 28 | - | 0.629133 | TAAAACGTTATGGCGTCT |
| chr2 | 11294476 | 11294625 | + | 11294491 | 11294507 | -60 | -45 | + | 0.628552 | GCGGAGGTGATGGGCCG |
| chr22 | 37215876 | 37216025 | + | 37215949 | 37215965 | -2 | 13 | + | 0.628528 | AGGTAGACATAGCTGCT |
| chr14 | 23290276 | 23290425 | + | 23290300 | 23290316 | -51 | -36 | - | 0.628377 | CAGAAATCATAGTACAC |
| chr14 | 23290226 | 23290375 | + | 23290300 | 23290316 | -1 | 14 | - | 0.628377 | CAGAAATCATAGTACAC |
| chr3 | 126080226 | 126080375 | + | 126080311 | 126080327 | 10 | 25 | + | 0.628183 | AGGAGCAGGAAGAGACC |
| chr3 | 169540276 | 169540425 | + | 169540277 | 169540293 | -74 | -59 | + | 0.62697 | TGGAAATCATTGACCTG |
| chr7 | 29186226 | 29186375 | + | 29186238 | 29186254 | -63 | -48 | - | 0.626717 | CCCAACCTCCTGGTCCC |
| chr10 | 130959576 | 130959725 | + | 130959675 | 130959691 | 24 | 39 | + | 0.626176 | AGCACCCACGTGTAACA |
| chr2 | 202753176 | 202753325 | + | 202753243 | 202753259 | -8 | 7 | + | 0.625963 | TGGAAGGTTTTGCATTT |
| chr14 | 106095376 | 106095525 | + | 106095480 | 106095496 | 29 | 44 | - | 0.624966 | GAGCTCAGGTCGGTCGT |
| chr7 | 2728726 | 2728875 | + | 2728847 | 2728863 | 46 | 61 | + | 0.624634 | CAGAGCATGGTTTCTTG |
| chr22 | 19743976 | 19744125 | + | 19744079 | 19744095 | 28 | 43 | + | 0.624388 | GAGGAGCAGATGTCTCA |
| chr13 | 28562426 | 28562575 | + | 28562456 | 28562472 | -45 | -30 | - | 0.624313 | GGGCTCAAGTGATCCTC |
| chr16 | 876076 | 876225 | + | 876175 | 876191 | 24 | 39 | - | 0.624069 | TACAACTCAGCGTTTGC |
| chr19 | 35818776 | 35818925 | + | 35818842 | 35818858 | -9 | 6 | - | 0.623746 | CAATGCTGTCTGTTCTG |
| chr19 | 35818726 | 35818875 | + | 35818842 | 35818858 | 41 | 56 | - | 0.623746 | CAATGCTGTCTGTTCTG |
| chr7 | 1747926 | 1748075 | + | 1747942 | 1747958 | -59 | -44 | + | 0.623692 | GAGAATCCTCAGCACCA |
| chr7 | 1747876 | 1748025 | + | 1747942 | 1747958 | -9 | 6 | + | 0.623692 | GAGAATCCTCAGCACCA |
| chr7 | 1329226 | 1329375 | + | 1329271 | 1329287 | -30 | -15 | + | 0.623404 | GGGCCCATTTCTTTTTT |
| chr10 | 13039576 | 13039725 | + | 13039636 | 13039652 | -15 | 0 | - | 0.622994 | CAGCACTTTTCCTTTCC |
| chr12 | 107297026 | 107297175 | + | 107297028 | 107297044 | -73 | -58 | + | 0.622992 | AGATCAAAAAAGTTCCT |
| chr12 | 107296976 | 107297125 | + | 107297028 | 107297044 | -23 | -8 | + | 0.622992 | AGATCAAAAAAGTTCCT |
| chr19 | 51538076 | 51538225 | + | 51538184 | 51538200 | 33 | 48 | - | 0.622895 | TCTACCACTCTGCACCT |
| chr17 | 81036076 | 81036225 | + | 81036193 | 81036209 | 42 | 57 | + | 0.622654 | AGACTCACAATTTTCAA |
| chr10 | 73324326 | 73324475 | + | 73324349 | 73324365 | -52 | -37 | + | 0.622338 | AGGACCATGCGGTATGG |
| chr3 | 129326576 | 129326725 | + | 129326702 | 129326718 | 51 | 66 | - | 0.622272 | GTGAGGACCATGCGCCG |
| chr17 | 7082976 | 7083125 | + | 7083005 | 7083021 | -46 | -31 | - | 0.622233 | CAGAGGGACCGGTTCTC |
| chr10 | 131650576 | 131650725 | + | 131650625 | 131650641 | -26 | -11 | - | 0.622231 | CAGTTCAACTTGGGATT |
| chr10 | 131650526 | 131650675 | + | 131650625 | 131650641 | 24 | 39 | - | 0.622231 | CAGTTCAACTTGGGATT |
| chr12 | 118312276 | 118312425 | + | 118312340 | 118312356 | -11 | 4 | - | 0.62212 | TTGTAAATATCATTCTA |
| chr4 | 1535626 | 1535775 | + | 1535695 | 1535711 | -6 | 9 | + | 0.62185 | AAGGTCACCGTCAACAG |
| chr4 | 1535576 | 1535725 | + | 1535695 | 1535711 | 44 | 59 | + | 0.62185 | AAGGTCACCGTCAACAG |
| chr20 | 36037676 | 36037825 | + | 36037686 | 36037702 | -65 | -50 | - | 0.621737 | AGCCTCAAGGCTTGCCT |
| chr1 | 3507026 | 3507175 | + | 3507140 | 3507156 | 39 | 54 | - | 0.621135 | GCCTACTCCTCGCTCCT |
| chr4 | 3690676 | 3690825 | + | 3690707 | 3690723 | -44 | -29 | - | 0.620909 | TTGTGAAAAATGATCCC |
| chr10 | 112834776 | 112834925 | + | 112834890 | 112834906 | 39 | 54 | + | 0.620836 | GAGCAGAAGACTTTAGTA |
| chr1 | 226791276 | 226791425 | + | 226791316 | 226791332 | -35 | -20 | + | 0.62068 | TGGGAAGTAGTGGCCAC |
| chr14 | 104768426 | 104768575 | + | 104768433 | 104768449 | -68 | -53 | - | 0.620472 | AAACACACCACGACTTT |
| chr14 | 104768376 | 104768525 | + | 104768433 | 104768449 | -18 | -3 | - | 0.620472 | AAACACACCACGACTTT |
| chr22 | 19743926 | 19744075 | + | 19744040 | 19744056 | 39 | 54 | + | 0.620272 | GAGAAATGGGCGTCTTG |
| chr19 | 49528526 | 49528675 | + | 49528537 | 49528553 | -64 | -49 | + | 0.619949 | AACCACATGTGGTGGCA |
| chr19 | 49528476 | 49528625 | + | 49528537 | 49528553 | -14 | 1 | + | 0.619949 | AACCACATGTGGTGGCA |
| chr15 | 68699576 | 68699725 | + | 68699654 | 68699670 | 3 | 18 | - | 0.619431 | AATTACAAAGTGCTGGT |
| chr15 | 68699526 | 68699675 | + | 68699654 | 68699670 | 53 | 68 | - | 0.619431 | AATTACAAAGTGCTGGT |
| chr6 | 130992626 | 130992775 | + | 130992704 | 130992720 | 3 | 18 | + | 0.619084 | GAGCAGACGAGTAACTT |
| chr1 | 9341926 | 9342075 | + | 9342044 | 9342060 | 43 | 58 | + | 0.618991 | ATGTCCAGGTCCTGCCA |
| chr10 | 118084276 | 118084425 | + | 118084361 | 118084377 | 10 | 25 | - | 0.618018 | GGCTCATCTTCTTGCT |
| chr1 | 226791326 | 226791475 | + | 226791436 | 226791452 | 35 | 50 | - | 0.6179 | AGCTACTGTCCGTCTTC |
| chr17 | 81036026 | 81036175 | + | 81036053 | 81036069 | -48 | -33 | - | 0.617564 | CTGAACAGATCCTTACA |
| chr21 | 46420426 | 46420575 | + | 46420429 | 46420445 | -72 | -57 | - | 0.617402 | GTGACCACCCCTTTGTC |
| chr10 | 131752776 | 131752925 | + | 131752899 | 131752915 | 48 | 63 | + | 0.617126 | GTGGGCCCATTGATCTC |
| chr1 | 55504626 | 55504775 | + | 55504737 | 55504753 | 36 | 51 | - | 0.616934 | AGGAGCAGCTAGTTGGT |
| chr10 | 4386676 | 4386825 | + | 4386729 | 4386745 | -22 | -7 | - | 0.616684 | TTCAACAGATCATTCCA |
| chr8 | 81963226 | 81963375 | + | 81963262 | 81963278 | -39 | -24 | + | 0.615925 | AGATAAAGCTTATACAT |
| chr14 | 101123426 | 101123575 | + | 101123558 | 101123574 | 57 | 72 | - | 0.615676 | AGGCAGGCTCTGGGTTC |
| chr22 | 37499376 | 37499525 | + | 37499427 | 37499443 | -24 | -9 | + | 0.614353 | AGCCACCTGGGGGGCCT |
| chr10 | 131650326 | 131650475 | + | 131650461 | 131650477 | 60 | 75 | + | 0.614225 | GGGAAGGATACATATTT |
| chr8 | 144367176 | 144367325 | + | 144367218 | 144367234 | -33 | -18 | - | 0.613893 | AGGTGCACTCGCACCCT |
| chr6 | 37527226 | 37527375 | + | 37527232 | 37527248 | -69 | -54 | + | 0.612857 | AGGCAGAGCTGGAGCTG |
| chr12 | 6933126 | 6933275 | + | 6933160 | 6933176 | -41 | -26 | - | 0.611805 | TGGATCATCCTCAGCAT |
| chr7 | 1328976 | 1329125 | + | 1329065 | 1329081 | 14 | 29 | - | 0.611154 | AGGGACACCAGCTCTAT |
| chr17 | 25798876 | 25799025 | + | 25798981 | 25798997 | 30 | 45 | - | 0.610396 | TGAAACTCAGTTCTCTG |
| chr5 | 137224876 | 137225025 | + | 137224895 | 137224911 | -56 | -41 | + | 0.610097 | GAGAAGCCCGTGATAAT |
| chr11 | 45392426 | 45392575 | + | 45392551 | 45392567 | 50 | 65 | - | 0.608851 | CAGTACAGTTATTGCGG |
| chr21 | 42219626 | 42219775 | + | 42219670 | 42219686 | -31 | -16 | + | 0.607964 | ATTTCCTGAATGTTCTG |
| chr1 | 234039826 | 234039975 | + | 234039923 | 234039939 | 22 | 37 | + | 0.607799 | AGGAAAAAGATGAGAAG |
| chr4 | 1160726 | 1160875 | + | 1160855 | 1160871 | 54 | 69 | - | 0.607587 | CAGCACAGCCTAGAGCA |
| chr2 | 66743676 | 66743825 | + | 66743807 | 66743823 | 56 | 71 | + | 0.606606 | GGTTGTAGAATCTTCCA |
| chr9 | 129387076 | 129387225 | + | 129387214 | 129387230 | 63 | 78 | - | 0.604888 | ATGAAAAGGATTTCGCC |
| chr5 | 137225126 | 137225275 | + | 137225236 | 137225252 | 35 | 50 | + | 0.604621 | AGGGACAGGGGGCGATTT |
| chr22 | 43829676 | 43829825 | + | 43829736 | 43829752 | -15 | 0 | - | 0.604372 | TGGCACCGCTTTGGTT |
| chr5 | 137225076 | 137225225 | + | 137225188 | 137225204 | 37 | 52 | - | 0.602888 | GGCGTCACGGTGGCACC |
| chr5 | 137224926 | 137225075 | + | 137224999 | 137225015 | -2 | 13 | + | 0.602796 | GGGCGAAAGGCCTCCCG |
| chr10 | 4378476 | 4378625 | + | 4378596 | 4378612 | 45 | 60 | + | 0.602635 | AAGTAAGATGCCTAACA |
| chr7 | 99067126 | 99067275 | + | 99067144 | 99067160 | -57 | -42 | + | 0.600841 | AGGCAGGTGAATTTTCC |
| chr4 | 1537176 | 1537325 | + | 1537305 | 1537321 | 54 | 69 | + | 0.599988 | GGAGCCGAGGAGATCCC |
| chr5 | 493176 | 493325 | + | 493182 | 493198 | -69 | -54 | + | 0.598359 | AGGCACCAACCTATCTG |
| chr14 | 104623476 | 104623625 | + | 104623611 | 104623627 | 60 | 75 | + | 0.594962 | CCGTACGTGGGGCAATT |
| chr12 | 52240326 | 52240475 | + | 52240392 | 52240408 | -9 | 6 | + | 0.593491 | AGTTACGGGATCCTCAG |
| chr1 | 218537326 | 218537475 | + | 218537398 | 218537414 | -3 | 12 | - | 0.593012 | AGACAATTCTTGTTTTT |
| chr5 | 162997876 | 162998025 | + | 162997958 | 162997974 | 7 | 22 | - | 0.591916 | AAATGCAAACCGTCCAA |
| chr19 | 1505826 | 1505975 | + | 1505941 | 1505957 | 40 | 55 | + | 0.591863 | GGAAATACGTTGACGTT |
| chr8 | 143261876 | 143262025 | + | 143261923 | 143261939 | -28 | -13 | + | 0.587376 | GAGCAATATTTGACCTA |
| chr5 | 137225026 | 137225175 | + | 137225069 | 137225085 | -32 | -17 | - | 0.584956 | AGGAGCTGGGGGCTCGC |
| chr13 | 28562476 | 28562625 | + | 28562480 | 28562496 | -71 | -56 | + | 0.583806 | AGCCTCCCAAAGTGCTG |
| chr6 | 25761526 | 25761675 | + | 25761627 | 25761643 | 26 | 41 | - | 0.56754 | CTGAAATTGATGCTTCT |

# 9 Abbreviations

| | |
|---|---|
| ADT | anti-androgen therapy |
| AR-MethSig | *AR* methylation signature |
| BAF | B-allele frequency |
| bp | base pair |
| BS | bisulfite |
| CASCADE | Cancer tissue Collection After Death |
| cfDNA | cell-free DNA |
| CT | computed tomography |
| ct-MethSig | circulating tumour methylation signature |
| CTC | circulating tumour cell |
| ctDNA | circulating tumour DNA |
| EED | Embryonic Ectoderm Development |
| FFPE | formalin fixed paraffin embedded |
| GMM | Gaussian Mixture Model |
| H3K27ME3 | Tri-methylation of lysine 27 on histone H3 protein subunit |
| HSPC | Hormone-sensitive prostate cancer |

| | |
|---|---|
| LASSO | least absolute shrinkage and selection operator |
| LHRH | luteinizing hormone releasing hormone |
| LP-WGBS | Low pass whole genome bisulfite sequencing |
| mCRPC | metastatic castration-resistant prostate cancer |
| MethSig | methylation Signature |
| MMR | mismatch repair |
| mPC | metastatic prostate cancer |
| MRD | minimal residual disease |
| MRI | magnetic resonance imaging |
| MSigDB | Molecular signature database |
| NGS | next-generation sequencing |
| OS | overall survival |
| PARP | poly (ADP-ribose) polymerases |
| PC | principal component |
| PCA | principal component analysis |
| PE | paired-end |
| PRC2 | polycomb repressor complex 2 |
| PSA | prostate-specific antigen |

| | |
|---|---|
| RECIST | Response Evaluation Criteria in Solid Tumors |
| RFC | random forest classifier |
| SNV | single nucleotide variant |
| SUZ12 | suppressor of zesta 12 |
| TRUS-biopsy | trans-rectal ultrasound-guided biopsy |
| WGBS | whole genome bisulfite sequencing |
| WGS | whole genome sequencing |